# A PREDICTIVE VIEW OF THE DETECTION AND CHARACTERIZATION OF INFLUENTIAL OBSERVATIONS IN REGRESSION ANALYSIS

by

Wesley Johnson and Seymour Geisser*
University of California at Davis
and
University of Minnesota

Technical Report No. 365

January, 1980

*Research supported in part by the University of Minnesota Graduate School and NIH Grant GM 25271.

## 0. INTRODUCTION

The problem of detecting influential subsets of data in the general regression setup has been considered by several individuals. In particular, Cook (1979), Bingham (1977), and Cook and Weisberg (1979) have considered the problem of detecting those subsets of data which are most influential in terms of their effect on the estimation of parameters in linear regression; or more generally, those subsets which most affect the estimation of specified linear combinations of parameters. As a special case of the above, they consider the effect of subsets on the mean regression line for specified independent variables.

Johnson and Geisser (1979) have also considered the above special case, but from a rather different perspective. Instead of determining "estimative" influence, they determine the influence that subsets have on the prediction of future values for fixed independent variables. In the Bayesian mode, assuming "vague" priors, the usual normality and independence assumptions for the general linear model, they determine predictive densities for future values: (Aitchison and Dunsmore (1975), Geisser (1965, 1971). These predictive densitites may be based on full data sets, and on subset deleted data sets. The predictive density based on subset deleted data set will differ from that based on the full data set, and the discrepancy between densities will often be reflected in a difference in location and shape. A very influential subset will cause these densities to be very discrepant. Kullback-Leibler divergences are used to measure these discrepancies and subsets of fixed size are ordered from least to most influential according to the magnitudes of these divergences. Kullback-Leibler divergences between predictive densities are defined to be predictive influence functions (PIF). Johnson and Geisser show that in large

samples, these PIF's break up into the sum of two components; the first reflecting the difference in point predictions, and the second, the difference in widths of standard predictive intervals. While it is convenient to have a function which incorporates location and scale discrepancies in this way, it is a considerable disadvantage that it is necessary to pre-specify independent variables.

In this paper, the problem of detecting influential subsets is again approached from the predictivist view. Dependence of PIF's on specification of independent variables is suppressed by a simple device.

Suppose a set of observations from a general linear model has been observed, and that the goal is to predict a future vector, with the same independent variables, and with the same values on these variables. Then predictive densities may be determined based on full and subset deleted samples, and PIF's may be defined as Kullbach-Leibler divergences between these densities. Ellipsoids in which future vectors are expected to lie will differ with respect to location and shape, and it will be seen that PIF's reflect this fact. Predictive influence, defined in this way, is essentially the "collective" predictive influence that subsets have on the prediction of those observations that have already been observed.

In section 1, the requisite predictive densities and Kullback-Leibler divergences will be derived; and predictive influence defined. PIF's are derived and studied in section 2. Section 3 is devoted to an example.

## 1. DEFINITIONS OF PREDICTIVE INFLUENCE

### (1.1) Introduction

The prediction problem and the problem of determining influential subsets are considered from the same point of view as in Johnson and

Geisser (1979). The essential difference is that, instead of determining the influence subsets have on the prediction of single future values, influences on some special future vectors will be considered.

In this section, predictive densities for future vectors in normal linear models with the usual independence assumptions and vague priors will be derived for the variance known and unknown cases. Kullback-Leibler divergences will be defined, derived, and studied for arbitrary multivariate normal densities. And finally, predictive influence functions will be defined, as Kullback-Leibler divergences between densities. The results of this section essentially generalize those of section 1 in Johnson and Geisser (1979).

## (1.2) Predictive Densities

Let $G$ be some arbitrary set, possibly $\phi$, and let $\underset{\sim}{\theta} \in R^k$ be unknown. Assume the existence of the family of probability densities

$$F = \{f_m(\cdot|x, \underset{\sim}{\theta}) \underset{\sim}{\theta} \in R^m, x \in G, m = 1, 2, \ldots\}$$

where each density, $f_m(\cdot|x, \underset{\sim}{\theta})$, has an $m$ dimensional argument. Let $\underset{\sim}{Y}$ be an $n$ dimensional random vector with density $f_n(\cdot|x, \underset{\sim}{\theta})$, and assume that $\underset{\sim}{Y} = \underset{\sim}{y}$ has been observed. Further, assume that it is of interest to predict a future $m$ dimensional random vector which is independent of $\underset{\sim}{Y}$, say $\underset{\sim}{Z}$, with density $f_m(\cdot|w, \underset{\sim}{\theta})$, $w \in G$; and assume the existence of a prior probability, possibly improper, $p(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}$, for the vector $\underset{\sim}{\theta}$. The posterior density, when it exists, is then defined to be

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) = \frac{f_n(\underset{\sim}{y}|x, \underset{\sim}{\theta}) p(\underset{\sim}{\theta})}{\int f_n(\underset{\sim}{y}|x, \underset{\sim}{\theta})\, p(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}} .$$

The predictive density of a future vector $\underset{\sim}{Z}$, given $w$, $x$, and $\underset{\sim}{y}$, may now be defined as

(1.2.1) $f_m(\underset{\sim}{z}|w,x,\underset{\sim}{y}) = \int f_m(\underset{\sim}{z}|w,\underset{\sim}{\theta})\ p(\underset{\sim}{\theta}|\underset{\sim}{y})\ d\underset{\sim}{\theta}$ ,

where the fact that the functional form of the density may depend on $n$ is suppressed. Note that the density is independent of the unknown parameter $\underset{\sim}{\theta}$ , and consequently may be used to make inferences about future values of $\underset{\sim}{Z}$ .

Consider now the usual linear model

(1.2.2) $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

where $X$ is a full rank $n \times p$ matrix of observed values, $\underset{\sim}{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, and $\underset{\sim}{\varepsilon}$ is an $n \times 1$ vector which has a $N_n[\underset{\sim}{0}, \theta I]$ distribution; $\theta$ known. Let $f(\cdot|X, \underset{\sim}{\beta}, \theta) = N_n[X\underset{\sim}{\beta}, \theta I]$, $G = \{X|X$ an $n \times p$ matrix of full rank, $n = 1, 2, \ldots,\}$ and define $F$ as above; $N_n[\ ,\ ]$ defines an $n$ dimensional multivariate normal density with given mean vector and covariance matrix. Consider the so-called "flat" improper prior

(1.2.3) $p(\underset{\sim}{\beta}|\theta)\ d\underset{\sim}{\beta} \propto d\underset{\sim}{\beta}$

and assume the goal is to predict a future vector from the model

(1.2.4) $\underset{\sim}{Z} = W\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}^*$ ;

where $W$ is an $m \times p$ matrix of known values and $\underset{\sim}{\varepsilon}^*$ is a $N[0, \theta I]$ random vector. Then, the predictive density of a future vector $\underset{\sim}{Z}$ , given an observed vector $\underset{\sim}{y}$ sampled from (1.2.2) is $N_m[W\hat{\underset{\sim}{\beta}},\ (I+W(X'X)^{-1}W')\theta]$, i.e.

(1.2.5) $f_m(\underset{\sim}{z}|W,X,\underset{\sim}{y}) \propto \exp[-\tfrac{1}{2}(\underset{\sim}{z}-W\hat{\underset{\sim}{\beta}})' \dfrac{(I+W(X'X)^{-1}W')^{-1}}{\theta} (\underset{\sim}{z} - W\hat{\underset{\sim}{\beta}})]$,

where $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$, the usual least squares estimate of $\underset{\sim}{\beta}$ .

If it is assumed that $\theta$ is unknown and that

$$(1.2.6) \qquad p(\underset{\sim}{\beta}, \theta)\, d\underset{\sim}{\beta}\, d\theta \propto \theta^{-1} d\underset{\sim}{\beta}\, d\theta$$

then the predictive density of a future vector is

$$St_m\left[n-p,\ W\underset{\sim}{\beta},\ (I + W(X'X)^{-1}W')s^2\right], \quad \text{i.e.}$$

$$(1.2.7) \qquad f_m(\underset{\sim}{z}|W,X,\underset{\sim}{y}) \propto \left[1 + (\underset{\sim}{z}-W\underset{\sim}{\hat{\beta}})' \frac{(I+W(X'X)^{-1}W')^{-1}}{(n-p)s^2} (\underset{\sim}{z}-W\underset{\sim}{\hat{\beta}})\right]^{-\frac{(n-p+1)}{2}},$$

where $s^2 = (\underset{\sim}{y} - X\underset{\sim}{\hat{\beta}})'(\underset{\sim}{y} - X\underset{\sim}{\hat{\beta}})/(n-p)$, the usual regression mean square error, and $St_m[\ ,\ ,\ ]$ denotes an $m$ dimensional multivariate Student density with specified degrees of freedom, location vector, and dispersion matrix.

It is not difficult to show in the above cases that the predictive densities converge almost surely to the sampling density of the future vector, a necessary criterion for the use of the prior distribution, c.f. Geisser (1971). Also, Murray (1977) has shown that the predictive density is the optimal sampling density estimate in the frequency sense, among those densities which are invariant with respect to translations and non-singular transformations, using the Kullback-Leibler measure of divergence.

(1.3.) Kullback-Leibler Divergences

Let $f_1$ and $f_2$ be generalized densities with respect to some measure $\nu$. Let $E_{f_i}$ denote the operator which takes expectation with respect to the density $f_i$, $i = 1, 2$, and define the directed divergences

$$(1.3.1) \qquad I(f_1, f_2) = E_{f_1} \ln(f_1/f_2)$$

$$(1.3.2) \qquad I(f_2, f_1) = E_{f_2} \ln(f_2/f_1)$$

and the divergence

$$(1.3.3) \qquad J(f_1, f_2) = I(f_1, f_2) + I(f_2, f_1) \ .$$

General properties of these functions are given in Kullback (1968).

Divergences will now be derived for the general multivariate normal case. Let $f_1 = N_m[\underset{\sim}{\mu}_1, \Sigma_1]$ and $f_2 = N_m[\underset{\sim}{\mu}_2, \Sigma_2]$, $\Sigma_1$ and $\Sigma_2$ p.d., $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2 \in R^m$, and define $\underset{\sim}{\mu}' = (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1)' \Sigma_1^{-\frac{1}{2}}$ and $\Sigma = \Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}$. Then

$$2\, I(f_1, f_2) = \underset{\sim}{\mu}' \Sigma^{-1} \underset{\sim}{\mu} + [\mathrm{tr}(\Sigma^{-1}) - \ln|\Sigma^{-1}| - m] \tag{1.3.4}$$

$$2\, I(f_2, f_1) = \underset{\sim}{\mu}' \underset{\sim}{\mu} + [\mathrm{tr}(\Sigma) - \ln|\Sigma| - m] \tag{1.3.5}$$

and

$$2\, J(f_1, f_2) = \underset{\sim}{\mu}' (I + \Sigma^{-1}) \underset{\sim}{\mu} + \sum_{i=1}^{m} \left[ \frac{(\sigma_{ii} - 1)^2}{\sigma_{ii}} + \frac{\delta_i^2}{(1-\delta_i^2)\sigma_{ii}} \right], \tag{1.3.6}$$

where $\sigma_{ii} = \{\Sigma\}_{ii}$, and $\Sigma$ is partitioned as $\Sigma = \left[\begin{array}{c|c} \sigma_{ii} & \underset{\sim}{\sigma}_{i2} \\ \hline \underset{\sim}{\sigma}_{2i} & \Sigma_{22} \end{array}\right]$, $\delta_i^2 = \underset{\sim}{\sigma}_{i2} \Sigma_{22}^{-1} \underset{\sim}{\sigma}_{2i} / \sigma_{ii}$, $i = 1, 2, \ldots, m$. These results are shown by standard calculations.

It is clear, as it was in the univariate case, c.f. Johnson and Geisser (1979), that divergences are partitioned into two components, the first now reflecting the difference in mean vectors relative to some covariance structure, and the second reflecting the difference in covariance structures. Defining $f_2^* = N_m[\underset{\sim}{\mu}_2, \Sigma_1]$ and $f_2^+ = N_m[\underset{\sim}{\mu}_1, \Sigma_2]$,

$$2I(f_2^*, f_1) = \underset{\sim}{\mu}' \underset{\sim}{\mu}, \quad 2I(f_2^+, f_1) = \mathrm{tr}(\Sigma) - \ln|\Sigma| - m. \tag{1.3.7}$$

Hence, if interest rests solely on the difference in mean vectors or covariance structures, it is enough to consider only the first or second component respectively.

Geometrically, the pairs, $(\underset{\sim}{\mu}_1, \Sigma_1)$ and $(\underset{\sim}{\mu}_2, \Sigma_2)$, define ellipsoids in m dimensional Euclidean space with centers $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$, and shapes determined by $\Sigma_1$ and $\Sigma_2$, respectively. Since covariance matrices may be thought of as inner products, the first component of each divergence is then the squared distance between the vectors $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$, relative to some

inner product. Different inner products will then attach more or less significance to differences in mean vectors. The second component of each divergence is independent of the mean vectors; so consider the ellipsoids with center zero defined by $\Sigma_1$ and $\Sigma_2$. Now there exists a set of $m$ vectors through the center of these ellipsoids which define a set of conjugate axes, Dempster (1969). This set of vectors may be taken to be the set of eigenvectors of $\Sigma$, or equivalently, the set of eigenvectors of $\Sigma_1$ relative to $\Sigma_2$. Denote this set of eigenvectors as $\Gamma = (\Gamma_1, \Gamma_2, \ldots, \Gamma_m)$, and denote the corresponsing matrix of eigenvalues as $\Lambda = \text{diag}\ \{\lambda_i\}_{i = 1, \ldots, m}$. Now $\lambda_i$ may be interpreted as the ratio of the squared length of the semi-axis of the ellipsoid traced out by $\Sigma_1$ in the direction of $\Gamma_i$, to the squared length of the semi-axis of the ellipsoid traced out by $\Sigma_2$, in the same direction. Hence, $\text{tr}(\Sigma)$ may be interpreted as the sum of these square ratios over all $m$ conjugate axes. Further, $|\Sigma|^{\frac{1}{2}}$ may be interpreted as the ratio of the volumes of these ellipsoids. Hence it is seen that if all ratios are equal to one, the component of divergence reflecting the difference in ellipsoid shape will be zero, as the ellipsoids will necessarily be the same. Recalling that $\Gamma' \Sigma \Gamma = \Lambda$, it is possible to get canonical expressions for divergences. For example, the expression (1.3.5) may be written as

$$2I(f_2, f_1) = \underset{\sim}{\mu}' \underset{\sim}{\mu} + \left[ \sum_{i=1}^{m} (\lambda_i - \ln(\lambda_i) - 1) \right] .$$

Assume now that $\Sigma_1 > \Sigma_2$. Then it follows that

(1.3.8) $\quad I(f_1, f_2) > I(f_2, f_1)$.

To see this note by (1.3.4) and (1.3.5), that

$$2I(f_1, f_2) - 2I(f_2, f_1) - (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1)' (\Sigma_2^{-1} - \Sigma_1^{-1})(\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1) +$$

$$\mathrm{tr}[\Sigma_1\Sigma_2^{-1} - \Sigma_2\Sigma_1^{-1}] - 2 \ln |\Sigma_1\Sigma_2^{-1}| .$$

Since $\Sigma_2^{-1} - \Sigma_1^{-1}$ is p.d., the first term is positive. Now let $\Omega = \Sigma_2\Sigma_1^{-1}$. Then $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the roots of $\Omega$ and the last two terms on the right reduce to

$$\sum_{i=1}^{m}[(\lambda_i)^{-1} - \lambda_i + 2 \ln(\lambda_i)]$$

which is positive if $\lambda_i < 1$ for all $i$. But

$$0 > \underset{\sim}{a}'(\Sigma_2 - \Sigma_1)\underset{\sim}{a} = \underset{\sim}{a}'\Sigma_1^{\frac{1}{2}}(\Sigma - I)\Sigma_1^{\frac{1}{2}}\underset{\sim}{a}$$

so the roots of $\Sigma$ are $< 1$; hence it is seen that, as in the univariate case, Johnson and Geisser (1979), differences in mean vectors and covariance matrices will get different weightings from different divergences.

(1.4) Predictive Influence

Consider the setup defined in section (1.2). It will be convenient to denote the predictive density of a future $\underset{\sim}{Z}$ given w, x, and $\underset{\sim}{y}$ as $f_m(\cdot|w, x, \underset{\sim}{y}) = f(\cdot)$. Now let $\underset{\sim}{y}' = (\underset{\sim}{y}_1', \underset{\sim}{y}_2')$ where $\underset{\sim}{y}_1$ has dimension $n_1$ and $\underset{\sim}{y}_2$ has dimension $n_2$, $n_1 + n_2 = n$; and define the predictive density of a future $\underset{\sim}{Z}$ given w, $x_1 \in G$, and $\underset{\sim}{y}_1$ as $f_m(\cdot|w, x_1, \underset{\sim}{y}_1) = f_{(2)}(\cdot)$, where the notation suggests that the subset $\underset{\sim}{y}_2$ has been deleted and that a reduction in $\underset{\sim}{y}$ may imply a reduction in x.

As in Johnson and Geisser (1979), it is of interest to measure the discrepancy between the densities f and $f_{(2)}$ for all possible $\binom{n}{n_2}$ subset deletions. Kullback-Leibler divergences are again candidates for measuring these discrepancies. They are respectively $I(f, f_{(2)})$, $I(f_{(2)}, f)$ and $J(f, f_{(2)})$.

In normal paradigms, it will usually be the case that predictive

dispersion matrices associated with $f_{(2)}$ will be larger, when averaged over the sampling distribution of the data, than those associated with $f$ ; c.f. Geisser (1971). By larger, it is meant of course that the difference in dispersion matrices is positive definite. Consequently, $I(f_{(2)}, f)$ will usually dominate $I(f, f_{(2)})$, due to (1.3.8), indicating that the former measure gives a higher weighting to mean vector and covariance differences than the latter.

## 2. A SPECIAL CASE OF MULTIVARIATE PREDICTIVE INFLUENCE IN NORMAL LINEAR MODELS

### (2.1) Introduction

Consider the general linear model which was defined in (1.2.2). Then, given the prior defined by (1.2.3) when $\theta$ is assumed known, and given the goal of predicting a future vector of observations from the model (1.2.4), it follows from (1.2.5) that predictive densities are multivariate normal. Hence, by (1.3.4), (1.3.5) and (1.3.6), it is possible to write down explicit representations of predictive influence. When $\theta$ is unknown, however, and the prior (1.2.6) is assumed, predictive distributions are not multivariate normal, but multivariate Student. Hence, as in Johnson and Geisser (1979) , it will be useful to approximate exact predictive influence functions by substituting appropriately scaled normal densities for Student densities in the definitions of predictive influence.

Now, whether one is trying to predict a single future observation, or a vector of observations, it is not always possible to specify $\underset{\sim}{w}$ or $W$ in advance. Nonetheless, it is still of great interest to determine an ordering for subsets in terms of their influence on future predictions. It is possible to accomplish this by determining the

collective predictive influence that subsets have on the prediction of observations that have already been observed. To be more explicit, assume that a vector $\underset{\sim}{y}$ of observations has been observed on the linear model defined by (1.2.2), and assume the goal of predicting a future vector from the same model. Then predictive densities will be normal or Student, as above, with $W = X$. These densities will have different mean vectors and covariance matrices, and predictive influence functions will reflect these facts. Specifically, when $\theta$ is known, predictive influence is partitionable into the sum of two components; the first reflecting the difference in mean vectors relative to some inner product, and the second reflecting the difference in covariance structures. Each component measures the collective influence that a deleted subset has on all $n$ dimensions. Equivalently, ellipsoids in which future vectors are expected to fall will differ with respect to location and shape. Deleted subsets will affect each dimension of the location vector as well as each dimension of the predictive ellipsoid. Components of predictive influence collectively measure these aspects.

When $\theta$ is unknown, approximate predictive influence functions will be determined as previously indicated. However, no attempt will be made to justify such approximations, as was done in Johnson and Geisser (1979). It is conjectured, however, that similar results obtain, up to the fact that convergence would now be of order $n^{-1}$ instead of $n^{-2}$. Further, it will not be possible to study the behavior of exact predictive influence functions, as was previously done, since these functions may not be resolved into simple functions of the data. It will, however, be possible to carefully study approximate predictive influence functions. Approximate PIF's will be denoted as $\hat{I}(\cdot,\cdot)$ and $\hat{J}(\cdot,\cdot)$.

In what follows, it will be seen that the first components of $I(f_{(2)}, f)$ and $\hat{I}(f_{(2)}, f)$ are proportional to Cook's distance in the known $\theta$ and unknown $\theta$ cases respectively, Cook(1977), Bingham (1977). The first components of the remaining PIF's will be different due to the different inner products involved. It will be shown that all PIF's are asymptotically equivalent, and in particular when $\theta$ is known, they are all asymptotically equivalent to Cook's distance. Hence for this case, Cook's distance is seen to be a first order approximation to $I(f_{(2)}, f)$ . However, when $\theta$ is unknown, asymptotic PI will be partitionable into the sum of components; the first component will be equivalent to Cook's distance, and the second component will be a function of Bingham's statistic $T_1^2$ , Bingham (1977); Johnson and Geisser (1979). Hence, it would appear that Cook's distance is adequate for prediction when $\theta$ is known and sample sizes are large or if it is only of interest to determine the effect subsets have on predictive mean vectors. However, when $\theta$ is unknown, it appears that more general statistics are appropriate if prediction is the goal.

(2.2) Preliminaries

Let $\underset{\sim}{Y}$ be defined as in (1.2.2) and let $\underset{\sim}{y}$ denote a realization of the vector $\underset{\sim}{Y}$. Partition $\underset{\sim}{Y}'$ and $\underset{\sim}{y}'$ as $(\underset{\sim 1}{Y}', \underset{\sim 2}{Y}')$ and $(\underset{\sim 1}{y}', \underset{\sim 2}{y}')$ respectively where $\underset{\sim i}{Y}' = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ and $\underset{\sim i}{y}' = (y_{i1}, y_{i2}, \ldots, y_{in_i})$ , $i = 1, 2$; $n_1+n_2 = n$. Partition $X'$ as $(X_1', X_2')$ , where $X_1$ is $n_1 \times p$; and $X_2$ is $n_2 \times p$; and partition $\underset{\sim}{\varepsilon}'$ as $(\underset{\sim 1}{\varepsilon}', \underset{\sim 2}{\varepsilon}')$ where $\varepsilon_i$ is of dimension $n_i$, $i = 1, 2$. Then $\underset{\sim i}{Y} = X_i \underset{\sim}{\beta} + \underset{\sim i}{\varepsilon}$ $i = 1, 2$. Suppose we focus on predicting an $n$ dimensional vector of observations at $W = X$ . Then, given a prior

probability element, predictive densities based on full and subset deleted samples are defined as in Section (1.4) as

$$f_n(\underset{\sim}{z}|X, X, \underset{\sim}{y}) = f(\underset{\sim}{z}) \ , \ f_n(\underset{\sim}{z}|X, X_1, \underset{\sim}{y}_1) = f_{(2)}(\underset{\sim}{z})$$

respectively. PIF's are defined as divergences between these densities. When $\theta$ is known, all three functions will be determined and studied. When $\theta$ is unknown, only two functions will be considered.

(2.3) Case 1; $\theta$ Known.

Assume the prior defined in (1.2.3) and define

$$S = X'X, \ S_1 = X_1'X_1, \ S_2 = X_2'X_2 \ , \ \hat{\underset{\sim}{\beta}} = S^{-1}X'\underset{\sim}{y}, \ \hat{\underset{\sim}{\beta}}_{(2)} = S_1^{-1}X_1'\underset{\sim}{y}_1$$

$$\hat{\underset{\sim}{y}} = X\hat{\underset{\sim}{\beta}}, \ \hat{\underset{\sim}{y}}_{(2)} = X\hat{\underset{\sim}{\beta}}_{(2)}, \ \hat{\underset{\sim}{y}}_2 = X_2\hat{\underset{\sim}{\beta}}, \ \hat{\underset{\sim}{y}}_{2(2)} = X_2\hat{\underset{\sim}{\beta}}_{(2)}$$

$$V_2 = X_2S^{-1}X_2', \ U_2 = X_2S_1^{-1}X_2', \ M = XS^{-1}X' \ , \ M_{(2)} = XS_1^{-1}X' \ .$$

These are all familiar expressions from linear models theory. For example, $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{\beta}}_{(2)}$ denote the usual least squares estimates of $\underset{\sim}{\beta}$ based on the full and deleted data sets respectively; $V_2$ and $U_2$ are proportional to sampling covariance matrices for predicted vectors at $X_2$; etc. Now when $n_2 = 1$ and, say, observation $i$ has been deleted, the subscript (2) above will be replaced by $(i)$, e.g., $V_2 = v_i = \underset{\sim}{x}_i S^{-1} \underset{\sim}{x}_i'$ . Now, by (1.2.5), the predictive densities $f$ and $f_{(2)}$ are respectively $N_n[\hat{\underset{\sim}{y}}, (I + M)\theta]$ and $N_n[\hat{\underset{\sim}{y}}_{(2)}, (I + M_{(2)})\theta]$ respectively.

Digressing for the moment, let $\xi$ be the $100(1 - \alpha)$th percentage point of a $\chi^2(n)$ random variable. Then it is evident that ellipsoids defined by

$$\{z \mid (z - \hat{y})'(I+M)^{-1} (z - \hat{y}) \leq \xi\}, \{z \mid (z - \hat{y}_{(2)})'(I+M_{(2)})^{-1}(z - \hat{y}_{(2)}) \leq \xi\}$$

define $1 - \alpha$ tolerance ellipsoids as well as $1 - \alpha$ predictive ellipsoids for a future vector from the model (1.2.2). Hence subset deletion will affect the location and shape of such ellipsoids; resulting in concern regardless of whether one's approach is Bayesian or classical.

In passing, note that

$$(I + M)(I + M_{(2)}) = I + 2M_{(2)} + M = (I + M_{(2)})(I + M)$$

since $M$ is idempotent. Hence, the principal axes of the above ellipsoids are aligned in the same directions, and the difference in shape of the above ellipsoids depends only on the fact that principal axes have different lengths.

Before going on to determine representations of P.I., it will be useful to state some propositions.

Proposition (2.3.1):

$$(2.3.1) \quad S^{-1} = S_1^{-1} - S_1^{-1}X_2'(I + U_2)^{-1}X_2S_1^{-1} .$$

$$(2.3.2) \quad S_1^{-1} = S^{-1} + S^{-1}X_2'(I - V_2)^{-1}X_2S^{-1}$$

$$(2.3.3) \quad V_2 = U_2(I + U_2)^{-1}$$

$$(2.3.4) \quad (I + U_2)^{-1} = (I - V_2)$$

$$(2.3.5) \quad (y_2 - \hat{y}_2) = (I + U_2)^{-1}(y_2 - \hat{y}_{2(2)})$$

$$(2.3.6) \quad \hat{\beta} = \hat{\beta}_{(2)} + S_1^{-1}X_2'(I + U_2)^{-1}(y_2 - \hat{y}_{2(2)})$$

Proof: These identities are collected in Bingham (1977). ■

Proposition (2.3.2): Let $X = (e_n, X^*)$, $X_i = (e_{n_i}, X_i^*)$ for $i = 1,2$, and let $x_{ij}$ be the jth row of $X_i^*$, $j = 1,\ldots,n_i$, $i = 1,2$. Then define

$$\bar{x} = \sum_{i=1}^{2}\sum_{j=1}^{n_i} x_{ij}/n, \quad \bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i, \quad \tilde{X} = X^* - e_n \bar{x}, \quad \tilde{X}_i = X_i^* - e_{n_i}\bar{x}_i$$

$$S_x = \tilde{X}'\tilde{X}/n, \quad S_x^{(i)} = \tilde{X}_i'\tilde{X}_i/n_i, \quad J_n = e_n e_n', \quad J_{n_i} = e_{n_i} e_{n_i}'$$

for $i = 1, 2$. Then

$$U_2 = n_1^{-1}[J_{n_2} + (X_2^* - e_{n_2}\bar{x}_1)S_x^{(1)^{-1}}(X_2^* - e_{n_2}\bar{x}_1)'] \tag{2.3.7}$$

$$V_2 = n^{-1}[J_2 + (X_2^* - e_{n_2}\bar{x})S_x^{-1}(X_2^* - e_{n_2}\bar{x})'] \tag{2.3.8}$$

Proof: These results are shown in Johnson and Geisser (1979). ■

In (1977), Cook proposed a statistic to "determine the degree of influence the i th data point has on the estimate of $\hat{\beta}$." Cook assumes $\theta$ is unknown. However, when $\theta$ is known, Cook's distance may be defined as in Cook (1977).

$$D_i = (\hat{\beta} - \hat{\beta}_{(i)})'\frac{S}{p\theta}(\hat{\beta} - \hat{\beta}_{(i)}) \qquad i = 1, \ldots, n. \tag{2.3.9}$$

Cook (1977) shows that

$$D_i = p^{-1}t_i^2 v_i/(i-v_i) \tag{2.3.10}$$

where $t_i^2 = (y_i - \hat{y}_i)^2/(1-v_i)\theta$. Further, Bingham (1977), proposes a class of statistics

$$D_Q^2 = (\hat{\beta} - \hat{\beta}_{(2)})'Q(\hat{\beta} - \hat{\beta}_{(2)}) \tag{2.3.11}$$

for Q p.s.d. . Then if $Q = S/p\theta$, $D_Q^2$ is the appropriate generalization of Cook's distance for the case when $n_2$ is arbitrary. Using Proposition (2.3.1), Bingham shows that

$$(2.3.12)\quad D_Q^2 = (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_2)' X_2 S_1^{-1} Q S_1^{-1} X_2' (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_2)$$

$$= (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_{2(2)})' X_2 S^{-1} Q S^{-1} X_2 (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_{(2)})$$

$$(2.3.13)\quad D_S^2 = (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_2)' (I - V_2)^{-1} V_2 (I - V_2)^{-1} (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_2)$$

$$= (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_{2(2)})' U_2 (I + U_2)^{-1} (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_{2(2)})$$

and he notes that

$$(2.3.14)\quad D_S^2 = (\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)})' (\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)}) \quad .$$

Therefore each $D_Q^2$ measures the collective distance between the deleted subset and regression lines which were computed from full or subset deleted data sets. Further, Cook's distance is seen to be the squared Euclidean distance between the vectors $\hat{\underset{\sim}{y}}$ and $\hat{\underset{\sim}{y}}_{(2)}$. In what follows, it will be shown that each of the leading terms of PIF's may be expressed as $D_Q^2$ for some $Q$.

It will be necessary to indicate another sequence of propositions before going on to derive PIF's. These propositions are proved in Appendix (4.1).

Proposition (2.3.3): Let $M$ and $M_{(2)}$ be defined as above. Then

$$(2.3.15)\quad |I + M| = 2^p$$

$$(2.3.16)\quad |I + M_{(2)}| = 2^p \, |I + \tfrac{1}{2} V_2 (I - V_2)^{-1}| \quad .$$

Proposition (2.3.4): Given the above definitions

$$(2.3.17)\quad (I + M_{(2)})^{-1} = I - \tfrac{1}{2}M - \tfrac{1}{2}XS^{-1}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2 S^{-1} X' .$$

Proposition (2.3.5):

$$(2.3.18)\quad \mathrm{tr}(I + M)^{-1}(I + M_{(2)}) = n + \tfrac{1}{2}\mathrm{tr}[V_2 (I - V_2)^{-1}]$$

(2.3.19) $\operatorname{tr}(I + M_{(2)})^{-1}(I + M) = n - \tfrac{1}{2}\operatorname{tr}[V_2(I - \tfrac{1}{2}V_2)^{-1}]$ .

Now define

$$Q_1 = S/p\theta \ , \quad Q_2 = [S - \tfrac{1}{2}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2]/p\theta \ , \quad Q_3 = Q_1 + Q_2 \ .$$

Then

(2.3.20) $$2I(f, f_{(2)}) = \frac{p}{2} D^2_{Q_2} + \ln|I + \tfrac{1}{2}V_2(I-V_2)^{-1}| - \tfrac{1}{2}\operatorname{tr}[V_2(I- \tfrac{1}{2}V_2)^{-1}] \ ,$$

(2.3.21) $$2I(f_{(2)},f) = \frac{p}{2} D^2_{Q_1} + \frac{1}{2}\operatorname{tr}[V_2(I - V_2)^{-1}] - \ln|I + \tfrac{1}{2}V_2(I - V_2)^{-1}|$$

(2.3.22) $$2J(f, f_{(2)}) = \frac{p}{2} D^2_{Q_3} + \tfrac{1}{2}\operatorname{tr}\{V_2[(I - V_2)^{-1} - (I - \tfrac{1}{2}V_2)^{-1}]\} \ .$$

First, consider the result (2.3.21). By (1.3.4),

$$2I(f_{(2)}, f) = (\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)})'\frac{(I + M)}{\theta}^{-1}(\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)}) +$$

$$[\operatorname{tr}(I + M)^{-1}(I + M_{(2)}) - \ln|(I+M)^{-1}(I+M_{(2)})| - n].$$

But $M$ is idempotent and $(I + M)^{-1} = (I - \tfrac{1}{2}M)$ . Hence, by Propositions (2.3.3) and (2.3.5), and the result (2.3.14), the above equals

$$\frac{1}{2\theta}(\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)})'(\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)}) + \tfrac{1}{2}\operatorname{tr}[V_2(I-V_2)^{-1}] - \ln|I+ \tfrac{1}{2}V_2(I-V_2)^{-1}|$$

$$= \frac{p}{2} D^2_{Q_1} + \tfrac{1}{2}\operatorname{tr}[V_2(I-V_2)^{-1} - \ln|I + \tfrac{1}{2}V_2(I-V_2)^{-1}|$$

and (2.3.21) is true. To verify the result (2.3.20), recall from (1.3.5) that

$$2I(f, f_{(2)}) = (\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)})'\frac{(I + M_{(2)})}{\theta}^{-1}(\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)}) +$$

$$[\operatorname{tr}(I + M)(I + M_{(2)})^{-1} - \ln|(I + M)(I + M_{(2)})^{-1}| - n].$$

Now by Propositions (2.3.3), (2.3.4) and (2.3.5), this equals

$$\frac{1}{\theta}\,(\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)})'(I - \tfrac{1}{2}M - \tfrac{1}{2}XS^{-1}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2S^{-1}X')(\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)}) +$$

$$\ln|I + \tfrac{1}{2}V_2(I-V_2)^{-1}| - \tfrac{1}{2}\,\mathrm{tr}[V_2(I - \tfrac{1}{2}V_2)^{-1}]$$

$$= \frac{1}{2\theta}\,(\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)})'(\hat{\underset{\sim}{y}} - \hat{\underset{\sim}{y}}_{(2)}) - \tfrac{1}{2}(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(2)})'X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(2)})$$

$$+ \ln|I + \tfrac{1}{2}V_2(I-V_2)^{-1}| - \tfrac{1}{2}\,\mathrm{tr}[V_2(I - \tfrac{1}{2}V_2)^{-1}]\,.$$

But by (2.3.11) and (2.3.14), the above becomes

$$\tfrac{1}{2}\,D_S^2 - \tfrac{1}{2}\,D^2_{X_2'(I - \frac{1}{2}V_2)^{-1}X_2} + \ln|I + \tfrac{1}{2}V_2(I-V_2)^{-1}| - \tfrac{1}{2}\,\mathrm{tr}[V_2(I - \tfrac{1}{2}V_2)^{-1}]$$

$$= \frac{p}{2}\,D^2_{Q_2} + \ln|I + \tfrac{1}{2}V_2(I-V_2)^{-1}| - \tfrac{1}{2}\,\mathrm{tr}[V_2(I - \tfrac{1}{2}V_2)^{-1}]$$

and (2.3.20) obtains. The result (2.3.22) follows by simple addition.

The above results clearly indicate that the first component of each PIF may, by (2.3.11), be interpreted as a measure of the distance between estimated regression coefficient vectors, relative to some inner product. Further, by (2.3.12) this component may be interpreted as the distance between $\underset{\sim}{y}_2$ and $\hat{\underset{\sim}{y}}_2$, relative to some inner product. Hence, the farther $\underset{\sim}{y}_2$ is collectively observed from the full, or deleted subset, regression lines, the greater will be the effect on predictive mean vectors. As previously indicated, this term is proportional to Cook's distance when $I(f_{(2)}, f)$ is considered.

The second component of P.I. reduces to a function of the matrix $V_2$ ; which, by (2.3.8), measures, in some sense, the distance between where the deleted subset is collectively observed, and where the center of the full data set lies. It is apparent that the farther $X_2^*$ is from $\underset{\sim}{e}_{n_2}\overline{\underset{\sim}{X}}$ , the greater will be the effect on predictive covariance matrices.

Now, it can easily be verified that $I + M_{(2)} \geq I + M$. By (2.3.1)

$$\underset{\sim}{t}'[(I + M_{(2)}) - (I + M)]\underset{\sim}{t} = \underset{\sim}{t}'X[S_1^{-1} - S^{-1}]X'\underset{\sim}{t}$$

$$= \underset{\sim}{t}'XS_1^{-1}X_2'(I + U_2)^{-1}X_2S_1^{-1}X'\underset{\sim}{t} \geq 0$$

as the matrix is of the form $\underset{\sim}{c}'\underset{\sim}{c}$ establishing the result. Then (1.3.8) implies that $I(f_{(2)}, f) \geq I(f, f_{(2)})$ under all circumstances. Hence the PIF $I(f_{(2)}, f)$ will always weight the differences in predictive mean vectors and covariance matrices more heavily than $I(f, f_{(2)})$.

Let $n_2 = 1$ and define

$$E_i = \frac{t_i^2}{p}\frac{v_i}{(1-v_i/2)}, \quad \alpha_i = \ln(1 + \tfrac{1}{2}\frac{v_i}{1-v_i}) - \frac{v_i}{2-v_i}$$

$$D_i = \frac{t_i^2}{p}\frac{v_i}{1-v_i}, \quad \lambda_i = \tfrac{1}{2}\frac{v_i}{1-v_i} - \ln(1 + \tfrac{1}{2}\frac{v_i}{1-v_i})$$

$$F_i = \frac{t_i^2}{p}\frac{v_i(4 - 3v_i)}{(2-v_i)(1-v_i)}, \quad \xi_i = \frac{\frac{1}{2}v_i^2}{(2-v_i)(1-v_i)}.$$

Then

$$(2.3.23)\quad 2I(f, f_{(i)}) = \frac{p}{2}E_i + \alpha_i, \quad 2I(f_{(i)}, f) = \frac{p}{2}D_i + \lambda_i, \quad 2J(f, f) = \frac{p}{2}F_i + \xi_i.$$

In this case then, it is seen that the first component of predictive influence is a weighted function of $t_i^2$; where the weight for each function is a monotone and increasing function of $v_i$. Hence, influence on predictive mean vectors is greatest when $y_i$ is distant from $\hat{y}_i$ and when $\underset{\sim}{x}_i^*$ is distant from $\bar{\underset{\sim}{x}}$. The second component of predictive influence is a monotone increasing function of $v_i$, so predictive covariance matrices are affected the most when $\underset{\sim}{x}_i^*$ is distant from $\bar{\underset{\sim}{x}}$. As in Johnson and Geisser (1979), different PIF's may order subsets differently due to the fact that both components of predictive influence have different weighting factors associated with them. Hence, one's choice of function could depend upon his convictions about the importance of these weighting factors.

Observe that $I(f_{(i)}, f)$ is proportional to the sum of Cook's distance, and a monotone increasing function of $v_i/(1-v_i)$; and that Cook's distance is the product of $t_i^2$ and $v_i/(1-v_i)$. It is possible to interpret the weighting function $v_i/(1-v_i)$. Recall from (2.3.2) that

$$u_j = \underset{\sim}{x}_j S_{(i)}^{-1} \underset{\sim}{x}_j' = v_j + v_{ji}^2/(1-v_i) .$$

Further, observe from (1.2.5) that the predictive variances for a single future value at $\underset{\sim}{w} = \underset{\sim}{x}_j$ for full and deleted subset data are respectively $\theta(1+v_j)$ and $\theta(1+u_j)$. Hence the sum of the differences of these variances is proportional to

$$\sum_{j=1}^{n} (u_j - v_j) = \sum_{j=1}^{n} v_{ji}^2/(1-v_i) = v_i/(1-v_i),$$

since $M$ is idempotent. Hence, $v_i/(1-v_i)$ is proportional to the sum of the extra variability due to the deletion of observation $i$, when predicting one at a time future values for the whole sample. Cook and Weisberg (1979) have equivalently shown that

$$\frac{v_i}{1-v_i} = \sum_{j=1}^{n} [\text{var}_{\text{samp.}}(\hat{y}_{j(i)}) - \text{var}_{\text{samp.}}(\hat{y}_j)]/\theta .$$

Consider a canonical representation of predictive influence. Recall the existence of an orthogonal matrix $\Gamma$ and a diagonal matrix $\Delta_2$ which satisfy $\Gamma V_2 \Gamma' = \Delta_2 \equiv \text{diag.}\ \{\Delta_i\}_{i=1}^{n_2}$ ; $\Gamma$ and $\Delta_2$ are the matrices of eigen-vectors and eigen-values of $V_2$ respectively. Now define

$$\underset{\sim}{r}_2 = (I-\Delta_2)^{-\frac{1}{2}} \Gamma(\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_2) / \theta \equiv (r_1, r_2, \ldots, r_{n_2})$$

$$\widetilde{E}_i = \frac{r_i^2}{p} \frac{\Delta_i}{1-\Delta_i/2}, \quad \widetilde{\alpha}_i = \ln(1 + \tfrac{1}{2} \frac{\Delta_i}{1-\Delta_i}) - \frac{\Delta_i}{2-\Delta_i},$$

$$\widetilde{D}_i = \frac{r_i^2}{p} \frac{\Delta_i}{1-\Delta_i}, \quad \widetilde{\lambda}_i = \tfrac{1}{2} \frac{\Delta_i}{1-\Delta_i} - \ln(1 + \tfrac{1}{2} \frac{\Delta_i}{1-\Delta_i})$$

$$\widetilde{F}_i = \frac{r_i^2}{p} \frac{\Delta_i(4-3\Delta_i)}{(2-\Delta_i)(1-\Delta_i)}, \quad \widetilde{\xi}_i = \frac{\frac{1}{2}\Delta_i^2}{(2-\Delta_i)(1-\Delta_i)} .$$

Then

$$(2.3.24)\quad 2I(f, f_{(2)}) = \sum_{j=n_1+1}^{n} (\frac{p}{2}\tilde{E}_j + \tilde{\alpha}_j) \quad , \quad 2I(f_{(2)}, f) = \sum_{j=n_1+1}^{n} (\frac{p}{2}\tilde{D}_j + \tilde{\lambda}_j)$$

$$2J(f, f_{(2)}) = \sum_{j=n_1+1}^{n} (\frac{p}{2}\tilde{F}_j + \tilde{\xi}_j).$$

These results are easily shown using (2.3.20), (2.3.21) and (2.3.22) and the fact that $\Gamma'\Gamma = I$. Considering these results, it is seen that PIF's are additive in canonical variates. And for each $j = n_1+1, \ldots, n$, the canonical variate $\frac{p}{2}\tilde{E}_j + \tilde{\alpha}_j$ may be interpreted in much the same way as $\frac{p}{2}E_j + \alpha_j$ was in the $n_2 = 1$ case; and similarly for other definitions of predictive influence.

We now consider asymptotic representations of PIF's. Recall the definitions of Proposition (2.3.2) and define

$$\underset{\sim}{s}_{xy}^{(i)} = (\frac{1}{n_i})\tilde{X}_i' y_i \qquad i = 1, 2 \qquad \hat{\underset{\sim}{\beta}}^{(1)} = S_x^{(1)^{-1}} \underset{\sim}{s}_{xy}^{(1)}$$

$$\hat{\underset{\sim}{\beta}}^{(2)} = S_x^{(2)^{-1}} \underset{\sim}{s}_{xy}^{(2)} \quad , \qquad \hat{\underset{\sim}{\beta}}^{(2,1)} = \hat{\underset{\sim}{\beta}}^{(2)} - \hat{\underset{\sim}{\beta}}^{(1)} \qquad n_2 > p$$

$$\bar{\underset{\sim}{X}}_{21} = \bar{\underset{\sim}{X}}_2 - \bar{\underset{\sim}{X}}_1 \quad , \qquad \hat{\delta} = \bar{y}_2 - \bar{y}_1 - \bar{\underset{\sim}{X}}_{21} \hat{\underset{\sim}{\beta}}^{(1)} \quad .$$

Of course, if $n_2 = 1$, $S_x^{(2)}$ and $\underset{\sim}{s}_{xy}^{(2)}$ are zero matrices. Now by (2.3.3) and (2.3.13), $D_S^2 = (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_{2(2)})' V_2 (\underset{\sim}{y}_2 - \hat{\underset{\sim}{y}}_{2(2)})$. And, by (2.3.8) and the fact that

$$(2.3.25)\quad \hat{\underset{\sim}{y}}_{2(2)} = \underset{\sim}{e}_{n_2} \bar{y}_1 + \bar{\underset{\sim}{X}}_2 \hat{\underset{\sim}{\beta}}^{(1)} + \underset{\sim}{e}_{n_2} \bar{\underset{\sim}{X}}_{21} \hat{\beta}^{(1)}$$

it follows from some algebra that

$$(2.3.26)\quad D_S^2 = \frac{n_2^2}{n}[\hat{\delta}^2 + (\underset{\sim}{s}_{yx}^{(2)} - \hat{\underset{\sim}{\beta}}^{(1)'} S_x^{(2)} + \frac{n_1}{n}\hat{\delta}\bar{\underset{\sim}{X}}_{21}) S_x^{-1} (\underset{\sim}{s}_{xy}^{(2)} - S_x^{(2)} \hat{\underset{\sim}{\beta}}^{(1)} + \frac{n_1}{n}\bar{\underset{\sim}{X}}_{21}\hat{\delta})] \quad .$$

When $n_2 = 1$, it follows that

$$(2.3.27)\quad D_i = \frac{1}{np\theta}[y_i-\bar{y}_1-(\underset{\sim}{x}_i^*-\underset{\sim}{\bar{X}}_1)\hat{\underset{\sim}{\beta}}^{(1)}]^2[1+\frac{(n-1)^2}{n^2}(\underset{\sim}{X}_i^*-\underset{\sim}{\bar{X}}_1)S_x^{-1}(\underset{\sim}{X}_i^*-\underset{\sim}{\bar{X}}_1)'],$$

and if $n_2 > p$

$$(2.3.28)\quad D_S^2 = \frac{n_2^2}{n}[\hat{\delta}^2+(\hat{\underset{\sim}{\beta}}^{(2,1)'}S_x^{(2)}+\frac{n_1}{n}\hat{\delta}\underset{\sim}{\bar{X}}_{21})S_x^{-1}(S_x^{(2)}\hat{\underset{\sim}{\beta}}^{(2,1)}+\frac{n_1}{n}\hat{\delta}\underset{\sim}{\bar{X}}_{21}')].$$

Hence, when $n_2 > p$, it is seen that Cook's distance function will be zero if $\hat{\delta} = 0$ and $\hat{\underset{\sim}{\beta}}^{(2,1)} = \underset{\sim}{0}$; i.e. if the point $(\underset{\sim}{\bar{X}}_2, \bar{y}_2)$ lies on the deleted subset regression line and if the slopes for deleted and non-deleted subset regression lines are the same. This is equivalent to saying that such regression lines are identical. For the $n_2 = 1$ case, it is sufficient that the deleted observation lie on the regression line for the rest of the data. Cook's distance functions will be large, on the other hand, when $(\underset{\sim}{\bar{X}}_2, \bar{y}_2)$ is distant from the deleted subset regression line, and if slopes are much different and also if centers for deleted and non-deleted subsets are much different.

Now assume that

$$S_x^{(1)} \to \Sigma_x^{(1)} \text{ p.d.}, \quad \underset{\sim}{s}_{xy}^{(1)} \to \underset{\sim}{\sigma}_{xy}^{(1)}, \quad \bar{y}_1 \to \mu_y^{(1)} \text{ a.s.}, \quad \underset{\sim}{\bar{X}}_1 \to \underset{\sim}{\mu}_x^{(1)}$$

as $n \to \infty$, and define

$$\underset{\sim}{\beta}^{(1)} = \underset{\sim}{\Sigma}_x^{(1)^{-1}}\underset{\sim}{\sigma}_{xy}^{(1)}, \qquad \tilde{\delta} = \bar{y}_2 - \mu_y^{(1)} - (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\mu}_x^{(1)})\underset{\sim}{\beta}^{(1)}.$$

Then it is shown in Appendix (4.2) that

$$(2.3.29)\quad 2nI(f_{(2)}, f) \approx 2nI(f, f_{(2)}) \approx nJ(f, f_{(2)}) \approx$$

$$\frac{n_2^2}{2\theta}[\tilde{\delta}^2+(\underset{\sim}{s}_{xy}^{(2)}-S_x^{(2)}\underset{\sim}{\beta}^{(1)}+\tilde{\delta}(\underset{\sim}{\bar{X}}_2-\underset{\sim}{\mu}_x^{(1)}))^T\Sigma_x^{(1)^{-1}}(\underset{\sim}{s}_{xy}^{(2)}-S_x^{(2)}\underset{\sim}{\beta}^{(1)}+\tilde{\delta}(\underset{\sim}{\bar{X}}_2-\underset{\sim}{\mu}_x^{(1)}))']$$

Further, when $n_2 = 1$, the above become

$$(2.3.30)\quad \frac{1}{2\theta}[y_i-\mu_y^{(1)}-(\underset{\sim}{x}_i^*-\underset{\sim}{\mu}_x^{(1)})\underset{\sim}{\beta}^{(1)}]^2[1+(\underset{\sim}{x}_i^*-\underset{\sim}{\mu}_x^{(1)})\Sigma_x^{(1)-1}(\underset{\sim}{x}_i^*-\underset{\sim}{\mu}_x^{(1)})'] + o(1) \quad \text{a.s.}$$

Thus, all representations of asymptotic predictive influence are equivalent to Cook's statistic. The effect of subset deletion on predictive covariance matrices is of smaller order than the effect on predictive mean vectors, and hence has a relatively diminished effect in large samples.

(2.4) Case 2; $\theta$ Unknown

Let $p(\underset{\sim}{\beta},\theta)$ denote the usual non-informative prior density for $(\underset{\sim}{\beta},\theta)$ which is defined in (1.2.6). Define

$$s^2 = (\underset{\sim}{y} - \hat{\underset{\sim}{y}})'(\underset{\sim}{y} - \hat{\underset{\sim}{y}}) / (n - p), \quad s^2_{(2)} = (\underset{\sim}{y}_1 - \hat{\underset{\sim}{y}}_{1(2)})'(\underset{\sim}{y}_1 - \hat{\underset{\sim}{y}}_{1(2)})'/(n_1 - p),$$

where $\hat{\underset{\sim}{y}}_{1(2)}$ is defined in the obvious way. Then it follows from (1.2.7) that f and $f_{(2)}$ are respectively $St_n[n-p, \hat{\underset{\sim}{y}}, (I+M)s^2]$ and $St_n[n_1-p, \underset{\sim}{y}_{(2)}, (I+M_{(2)})s^2_{(2)}]$.

It is now possible to write down expressions for the exact predictive influence. However, these functions may not be reduced to simple functions of the data. Consequently, little insight may be gleaned from determining these expressions. It will, however, be fruitful to consider approximate PIF's, where appropriately scaled multivariate normal densities are substituted for Student densities in the definitions of predictive influence. These functions simplify suitably, and may be studied carefully.

Since exact PIF's do not readily lend themselves to careful study, no attempt will be made to determine the convergence properties of these functions. It is conjectured, however, that approximate and exact PIF's will converge to one another, and that the order of convergence is $n^{-1}$.

Now, let $\underset{\sim}{Z}^*$ be a $St_m[k, \underset{\sim}{\mu}, \Sigma]$ density for $\Sigma$ p.d., m,k = 1,2, ..., and $\underset{\sim}{\mu} \in R^k$. Then Cornish (1954) essentially showed that $\text{cov}(\underset{\sim}{Z}^*) = \frac{k}{k-2}\Sigma$. Hence, it appears reasonable to approximate f and $f_{(2)}$ by $N_n[\hat{\underset{\sim}{y}}, (I+M)(\frac{n-p}{n-p-2})s^2]$ and $N_n[\hat{\underset{\sim}{y}}_{(2)}, (I+M_{(2)})(\frac{n_1-p}{n_1-p-2})s^2_{(2)}]$ densities respectively; $n \geq p + 3$. Let $\tilde{f}$ and $\tilde{f}_{(2)}$ denote these densities, and define the approximate predictive influence as $\hat{I}(f, f_{(2)}) = I(\tilde{f}, \tilde{f}_{(2)})$, $\hat{I}(f_{(2)}, f) = I(\tilde{f}_{(2)}, \tilde{f})$, and $\hat{J}(f, f_{(2)}) = J(\tilde{f}, \tilde{f}_{(2)})$.

It will be necessary to indicate some preliminary results before going on to derive approximate PIF's.

Proposition (2.4.1): Given the above notation,

$$(2.4.1) \qquad s^2 = \left[(n_1-p)s^2_{(2)} + (\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_{2(2)})'(I+U_2)^{-1}(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_{2(2)})\right]/(n-p)$$

$$(2.4.2) \qquad s^2_{(2)} = \left[(n-p)s^2 - (\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_2)'(I-V_2)^{-1}(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_2)\right]/(n_1-p) .$$

Proof: This result is shown in Bingham (1977). ■

Proposition (2.4.2): As in Bingham (1977), define

$$T_1^2 = \frac{1}{n_2}(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_{2(2)})'(I+U_2)^{-1}(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_{2(2)})/s^2_{(2)}$$

$$t_1^2 = \frac{1}{n_2}(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_2)'(I-V_2)^{-1}(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_2)/s^2 .$$

Then

$$(2.4.3) \qquad s^2/s^2_{(2)} = \left(\frac{n_1-p}{n-p}\right) + \left(\frac{n_2}{n-p}\right)T_1^2 , \quad s^2_{(2)}/s^2 = \left(\frac{n-p}{n_1-p}\right) - \left(\frac{n_2}{n_1-p}\right)t_1^2$$

$$(2.4.4) \qquad t_1^2 = T_1^2 \Big/ \left\{\left(\frac{n_1-p}{n-p}\right)\left[1 + \frac{n_2}{n_1-p}T_1^2\right]\right\} .$$

Further, let $n_2 > p$ and define

$$M_2 = X_2(X_2'X_2)^{-1}X_2' \,,\quad \underset{\sim}{\hat{y}}_{2(1)} = M_2\underset{\sim}{y}_2 \,,\quad s_2^2 = (\underset{\sim}{y}_2-\underset{\sim}{\hat{y}}_{2(1)})'(\underset{\sim}{y}_2-\underset{\sim}{\hat{y}}_{2(1)})/(n_2-p)$$

$$T_2^2 = \frac{n_1n_2}{n}[\hat{\delta}(1-\frac{n_1n_2}{n^2}\,\bar{\underset{\sim}{x}}_{21}S_x^{-1}\bar{\underset{\sim}{x}}_{21}')-\frac{n_2}{n}\,\bar{\underset{\sim}{x}}_{21}S_x^{-1}S_x^{(2)}\hat{\underset{\sim}{\beta}}^{(2,1)}]^2/s_{(2)}^2(1-\frac{n_1n_2}{n^2}\,\bar{\underset{\sim}{x}}_{21}S_x^{-1}\bar{\underset{\sim}{x}}_{21}')$$

$$V_2^2 = \frac{s_2^2}{s_{(2)}^2} \,,\quad W_2^2 = \hat{\underset{\sim}{\beta}}^{(2,1)'}(S_x^{(2)^{-1}} + S_x^{(1)^{-1}})^{-1}\,\hat{\underset{\sim}{\beta}}^{(2,1)}/(p-1)s_{(2)}^2.$$

Then

$$(2.4.5)\qquad T_1^2 = \frac{1}{n_2}\,T_2^2 + (1-\frac{p}{n_2})V_2^2 + (\frac{p}{n_2}-\frac{1}{n_2})W_2^2.$$

Proof: The result (2.4.3) is obvious from Proposition (2.4.1), and the result (2.4.4) is obvious from (2.4.3). The result (2.4.5) is shown in Johnson and Geisser (1979). ■ Note here that the ratio of mean square errors for full and deleted subset data is a monotone function of Bingham's statistic, $T_1^2$. Further note by (2.4.5) that this statistic is a convex function of the statistics $T_2^2$, $V_2^2$, and $W_2^2$, which have sampling distributions which are $F(1, n_1 - p)$, $F(n_2 - p, n_1 - p)$ and $F(p - 1, n_1 - p)$ respectively. These statistics measure respectively, the collective distance that $\underset{\sim}{y}_2$ is observed from $\underset{\sim}{\hat{y}}_{2(2)}$, relative to some metric, Johnson and Geisser (1979); the relative amount of scatter of $\underset{\sim}{y}_2$ about $\underset{\sim}{\hat{y}}_{2(1)}$ to that of $\underset{\sim}{y}_1$ about $\underset{\sim}{\hat{y}}_{1(2)}$; and the difference in the slopes of least squares regression lines for $\underset{\sim}{y}_1$ and $\underset{\sim}{y}_2$ respectively.

Now consider $\hat{I}(f, f_{(2)})$ and recall from (1.3.5) that this is proportional to

$$(2.4.6)\qquad (\underset{\sim}{\hat{y}} - \underset{\sim}{\hat{y}}_{(2)})'\,\frac{(I + M_{(2)})^{-1}}{s^2}(\underset{\sim}{\hat{y}} - \underset{\sim}{\hat{y}}_{(2)})\cdot(\frac{n_1-p-2}{n_1-p}) +$$

$$\{\mathrm{tr}\,[(I + M)(I + M_{(2)})^{-1}\,\frac{s^2}{s_{(2)}^2}(\frac{n-p}{n-p-2})(\frac{n_1-p-2}{n_1-p})] -$$

$$\ln|(I + M)(I + M_{(2)})^{-1}\,\frac{s^2}{s_{(2)}^2}\,(\frac{n-p}{n-p-2})(\frac{n_1-p-2}{n_1-p})| - n\} \quad .$$

At a glance, it is apparent that nearly all calculations necessary to simplify expression (2.4.6) have already been performed. Define

$$Q_1^* = \frac{\theta}{s^2} Q_1 \cdot \left(\frac{n-p-2}{n-p}\right) \quad , \quad Q_2^* = \frac{\theta}{s^2_{(2)}} Q_2 \cdot \left(\frac{n_1-p-2}{n_1-p}\right)$$

and recall from Proposition (2.4.2) that $s^2/s^2_{(2)} = \left(\frac{n_1-p}{n-p}\right)\left(1 + \frac{n_2}{n_1-p} T_1^2\right)$. Then by the proof of (2.3.20), expression (2.4.6) is equivalent to

$$(2.4.7)\quad 2\hat{I}(f, f_{(2)}) = \frac{p}{2} D^2_{Q_2^*} - \frac{1}{2}\left(\frac{n_1-p-2}{n-p-2}\right)\left(\frac{n_2}{n_1-p}\right) T_1^2 \operatorname{tr}[V_2(I - \tfrac{1}{2} V_2)^{-1}]$$

$$\ln\left|I + \tfrac{1}{2} V_2 (I-V_2)^{-1}\right| - \frac{1}{2}\left(\frac{n_1-p-2}{n-p-2}\right) \operatorname{tr}[V_2(I - \tfrac{1}{2} V_2)^{-1}]$$

$$+ n\left[\left(\frac{n_1-p-2}{n-p-2}\right)\left(1 + \frac{n_2}{n_1-p} T_1^2\right) - \ln\left(\frac{n_1-p-2}{n-p-2}\right)\left(1 + \frac{n_2}{n_1-p} T_1^2\right) - 1\right] .$$

Similarly,

$$(2.4.8)\quad 2\hat{I}(f_{(2)}, f) = \frac{p}{2} D^2_{Q_1^*} - \frac{1}{2}\left(\frac{n-p-2}{n_1-p-2}\right)\left(\frac{n_2}{n-p}\right) t_1^2 \operatorname{tr}[V_2(I-V_2)^{-1}]$$

$$\frac{1}{2}\left(\frac{n-p-2}{n_1-p-2}\right) \operatorname{tr}[V_2(I-V_2)^{-1}] - \ln\left|I + \tfrac{1}{2} V_2 (I-V_2)^{-1}\right|$$

$$+ n\left[\left(\frac{n-p-2}{n_1-p-2}\right)\left(1 - \frac{n_2}{n-p} t_1^2\right) - \ln\left(\frac{n-p-2}{n_1-p-2}\right)\left(1 - \frac{n_2}{n-p} t_1^2\right) - 1\right]$$

$\hat{J}(f, f_{(2)})$ may be similarly expressed.

Note that the first term of expression (2.4.6 ) is proportional to Cook's distance for the arbitrary $n_2$, $\theta$ unknown case; Bingham (1977). The second term of (2.4.6 ) is more complicated than in the $\theta$ known case, since it now depends on the data $\underset{\sim}{y}$ as well as the X matrix. This measure of the difference in covariance structures depends on the data through the variable $t_1^2$, and as was previously the case, it depends on the X matrix through the matrix $V_2$. The variables $D^2_{Q_1^*}$ and $t_1^2$ measure the distance between $\underset{\sim}{y}_2$ and $\hat{\underset{\sim}{y}}_{2(2)}$, relative to different metrics; and $V_2$ again measures, in some sense, the distance between

$X_2^*$ and $\underset{\sim}{e}_{n_2}\underset{\sim}{\bar{X}}$. Hence, it is conjectured that the most influential subsets will have large $t_1^2$ and $D^2_{Q_1^*}$ values, as well as a large value for the trace of $V_2$ and/or a large value for the determinant of $I+\frac{1}{2}V_2(I-V_2)^{-1}$. The essential difference between this case and the $\theta$ known case is due to the fact that the difference in covariance structures depends on the data through the variable $t_1^2$. In what follows, it will be shown that this component is of the same order as the first component. The situation is similar for the other definitions of PI.

Let $n_2 = I$ and define

$$T_i^2 = (y_i-\hat{y}_{i(i)})^2/(1+u_i)s^2_{(i)}, \quad t_i^2 = (y_i-\hat{y}_i)^2/(1-v_i)s^2 ,$$

$$D_i = p^{-1}t_i^2 v_i/(1-v_i), \quad E_i = 2p^{-1}T_i^2 v_i/(2-v_i) .$$

Then

$$(2.4.9)\quad 2\hat{I}(f, f_{(i)}) = (\tfrac{n-p-3}{n-p-1})\tfrac{p}{2}E_i + n[(\tfrac{n-p-3}{n-p-2})(1+\tfrac{T_i^2}{n-p-1}) - \ln(\tfrac{n-p-3}{n-p-2})(1+\tfrac{T_i^2}{n-p-1}) - 1]$$
$$+ \ln(1+\tfrac{1}{2}\tfrac{v_i}{1-v_i}) - (\tfrac{n-p-3}{n-p-2})\tfrac{v_i}{2-v_i} + - (\tfrac{n-p-3}{n-p-2})(\tfrac{T_i^2}{n-p-1})(\tfrac{v_i}{2-v_i}) .$$

$$(2.4.10)\quad 2\hat{I}(f_{(i)}, f) = \tfrac{p}{2}(\tfrac{n-p-2}{n-p})D_i + n[(\tfrac{n-p-2}{n-p-3})(1-\tfrac{t_i^2}{n-p}) - \ln(\tfrac{n-p-2}{n-p-3})(1-\tfrac{t_i^2}{n-p}) - 1] .$$
$$+ \tfrac{1}{2}(\tfrac{n-p-2}{n-p-3})(\tfrac{v_i}{1-v_i}) - \ln(1+\tfrac{1}{2}\tfrac{v_i}{1-v_i}) + - \tfrac{1}{2}(\tfrac{n-p-2}{n-p-3})(\tfrac{t_i^2}{n-p})(\tfrac{v_i}{1-v_i}) .$$

Now it is not difficult to show that $\frac{\partial}{\partial v_i} I(f, f_{(i)}) > 0$ for all $T_i^2$ and that $\frac{\partial}{\partial T_i^2} I(f, f_{(i)}) > 0$ if and only if $T_i^2 > \left(\frac{n-p-1}{n-p-3}\right)\cdot\left(1+\frac{(n-p-3)^2}{n}\left(\frac{v_i}{2-v_i}\right)\right) / \left(1+\frac{(n-p-3)}{n}\left(\frac{v_i}{2-v_i}\right)\right)$

Hence, $I(f, f_{(i)})$ is a monotone increasing function of $v_i$ for all $T_i^2$, and if $\frac{v_i}{2-v_i} > \frac{n}{(n-p-3)2}$, $I(f, f_{(i)})$ is increasing in $T_i^2$.

Otherwise, $I(f, f_{(i)})$ is J shaped in $T_i^2$. Note that when $v_i/(2-v_i) = (2n-1)^{-1}$, i.e., when $(v_i/2-v_i)$ is minimum, the expression on the right of the last inequaltiy converges to $\frac{1}{2}$ as $n \to \infty$. Hence, asymptotically, $I(f, f_{(i)})$ will achieve its minimum as a function of $T_i^2$ at a value no larger than $\frac{1}{2}$. Results for expression (2.4.10) may similarly be indicated. It is clear that the most influential observation will be associated with large $T_i^2$ and $v_i$ values. The fact that these functions may be J shaped in $T_i^2$ does not seem to be of great significance. This fact will be better understood when asymptotic predictive influence functions are studied.

It is possible to write down the usual canonical representation of predictive influence where appropriate changes are made in the definition of $\underset{\sim}{r}_2$ to account for the fact that $\theta$ is unknown. Further, it is possible to study these as functions of $r_i^2$ and $\Delta_i$. In exactly the same way as was done in the $n_2 = 1$ case, it may be shown that predictive influence functions are monotone and increasing in each $\Delta_i$ and are J shaped in each $r_i^2$, $i = 1, \ldots, n_2$.

In order to obtain an asymptotic representation of predictive influence, it will be necessary to derive an explicit representation of $T_1^2$ in large samples. Assume the convergence assumptions of section (2.3) and recall that $s_{(2)}^2 \overset{a.s.}{\to} \theta$ as $n \to \infty$. Then (a.s.)

$$(2.4.11)\quad T_1^2 = \frac{1}{n_2\theta}[n_2\tilde{\delta}^2 + (\underset{\sim}{y}_2 - \bar{y}_2\underset{\sim}{e}_{n_2} - \tilde{X}_2\underset{\sim}{\beta}^{(1)})'(\underset{\sim}{y}_2 - \bar{y}_2\underset{\sim}{e}_{n_2} - \tilde{X}_2\underset{\sim}{\beta}^{(1)})] + o(1) .$$

If $n_2 > p$, define

$$\tilde{W}_2^2 = (\hat{\underset{\sim}{\beta}}^{(2)} - \underset{\sim}{\beta}^{(1)})'S_x^{(2)}(\hat{\underset{\sim}{\beta}}^{(2)} - \underset{\sim}{\beta}^{(1)}) \;/\; (p-1) \quad .$$

Then

$$(2.4.12)\quad T_1^2 = [\frac{1}{n_2}(\frac{n_2\tilde{\delta}^2}{\theta}) + (1 - \frac{p}{n_2})\frac{S_2^2}{\theta} + (\frac{p}{n_2} - \frac{1}{n_2})\frac{\tilde{W}_2^2}{\theta}] + o(1) \text{ a.s.}.$$

The result (2.4.12) follows directly from the definition of $T_2^2$, (2.4.5), and convergence assumptions, however it may also be derived from (2.4.11). To derive the result (2.4.11), observe that by (2.3.25)

$$(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_{2(2)}) = (\tilde{\delta}\underset{\sim}{e}_{n_2}+\underset{\sim}{y}_2-\bar{\underset{\sim}{y}}_2\underset{\sim}{e}_{n_2} - \tilde{\underset{\sim}{X}}_2\,\underset{\sim}{\beta}^{(1)}) + o(1) \quad \text{a.s.} .$$

Then by (2.3.4), (2.3.8), and some algebra,

$$T_1^2 = \frac{1}{n_2}(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_{2(2)})'(I+U_2)^{-1}(\underset{\sim}{y}_2-\hat{\underset{\sim}{y}}_{2(2)}) \,/\, s^2_{(2)}$$

$$= \frac{1}{n_2}(\tilde{\delta}\underset{\sim}{e}_{n_2}+\underset{\sim}{y}_2-\bar{\underset{\sim}{y}}_2\underset{\sim}{e}_{n_2}-\tilde{\underset{\sim}{X}}_2\underset{\sim}{\beta}^{(1)})'(\tilde{\delta}\underset{\sim}{e}_{n_2}+\underset{\sim}{y}_2-\bar{\underset{\sim}{y}}_2\underset{\sim}{e}_{n_2}-\tilde{\underset{\sim}{X}}_2\underset{\sim}{\beta}^{(1)}) \,/\, \theta + o(1) \quad \text{a.s.} .$$

$$= [\tilde{\delta}^2+(\underset{\sim}{y}_2-\bar{\underset{\sim}{y}}_2\underset{\sim}{e}_{n_2}-\tilde{\underset{\sim}{X}}_2\underset{\sim}{\beta}^{(1)})'(\underset{\sim}{y}_2-\bar{\underset{\sim}{y}}_2\underset{\sim}{e}_{n_2}-\tilde{\underset{\sim}{X}}_2\underset{\sim}{\beta}^{(1)})] \,/\, \theta + o(1) \quad \text{a.s.} .$$

Asymptotic representations are now easy to obtain. Let

$$\tilde{I}(f_{(2)}, f) = 2\lim_{n\to\infty} n\hat{I}(f_{(2)}, f) \text{ a.s.} , \quad \tilde{I}(f, f_{(2)}) = 2\lim_{n\to\infty} n\hat{I}(f, f_{(2)}) \text{ a.s.}$$

$$\tilde{J}(f, f_{(2)}) = \frac{1}{2}[\tilde{I}(f, f_{(2)}) + \tilde{I}(f_{(2)}, f)] \quad \text{a.s.} .$$

Then (a.s.)

$$(2.4.13)\quad \tilde{I}(f, f_{(2)}) = \frac{n_2^2}{2}(T_1^2 - 1)^2 +$$

$$\frac{n_2^2}{2\theta}[\tilde{\delta}^2+(\underset{\sim}{s}^{(2)}_{xy}-S_x^{(2)}\underset{\sim}{\beta}^{(1)})'+\tilde{\delta}(\bar{\underset{\sim}{X}}_2-\underset{\sim}{\mu}_x^{(1)})')\Sigma_x^{(1)^{-1}}(\underset{\sim}{s}^{(2)}_{xy}-S_x^{(2)}\underset{\sim}{\beta}^{(1)}+\tilde{\delta}(\bar{\underset{\sim}{X}}_2-\underset{\sim}{\mu}_x^{(1)}))]+ o(1)$$

$$(2.4.14)\quad \tilde{I}(f, f_{(2)}) \approx \tilde{I}(f_{(2)}, f) \approx \tilde{J}(f, f_{(2)}) .$$

When $n_2 = 1$, the result (2.4.13) reduces to

$$(2.4.15)\quad \tilde{I}(f, f_{(i)}) = \frac{1}{2}T_i^2[1+(\underset{\sim}{x}_i^*-\bar{\underset{\sim}{X}}_1)S_x^{(1)^{-1}}(\underset{\sim}{x}_i^*-\bar{\underset{\sim}{X}}_1)'] + \frac{1}{2}[T_i^2 - 1]^2 + o(1) \quad \text{a.s.}$$

$$= \frac{\tilde{\delta}^2}{2\theta}[1+(\underset{\sim}{x}_i^*-\underset{\sim}{\mu}_x^{(1)})\Sigma_x^{(1)^{-1}}(\underset{\sim}{x}_i^*-\underset{\sim}{\mu}_x^{(1)})'] + \frac{1}{2}[\frac{\tilde{\delta}^2}{\theta} - 1]^2+o(1) \text{ a.s.}$$

where

$$\tilde{\delta} = [y_i - \mu_y^{(1)} - (\underset{\sim}{x}_i^* - \underset{\sim}{\mu}_x^{(1)})\underset{\sim}{\beta}^{(1)}] \quad .$$

Consider the result (2.4.7). The first component clearly converges appropriately since $Q_1^* \xrightarrow{a.s.} Q_1$ as $n \to \infty$, and the result has already been shown for $D^2_{Q_1}$. The fact that

$$\ln|I + \frac{1}{2} V_2(I-V_2)^{-1}| - \frac{1}{2}(\frac{n_1-p-2}{n-p-2}) \operatorname{tr}[V_2(I - \frac{1}{2} V_2)^{-1}] = o(n^{-1})$$

is easily shown by modifying the proof of the comparable result in section (2.3). Since $T_1^2$ converges a.s., (2.3.8) implies that

$$-\frac{1}{2}(\frac{n_1-p-2}{n-p-2})(\frac{n_2}{n_1-p})T_1^2 \operatorname{tr}[V_2(I-V_2)^{-1}] = o(n^{-1}) \qquad \text{a.s.} \; .$$

Finally, it is easy to demonstrate by a Taylor expansion that

$$n^2[(\frac{n_1-p-2}{n-p-2})(1 + \frac{n_2}{n_1-p} T_1^2 - \ln(\frac{n_1-p-2}{n-p-2})(1 + \frac{n_2}{n_1-p} T_1^2) - 1] = \frac{n_2^2}{2}[T_1^2 - 1]^2 + o(1) \text{ a.s.};$$

hence (2.4.13) obtains. The result (2.4.14) is similarly derived.

From (2.4.13), it is seen that asymptotic predictive influence is partitioned into the sum of Cook's distance, and a convex function of Bingham's statistic. By considering expressions (2.3.28) and (2.4.12) it is possible to distinguish between Cook's and Bingham's statistics when $n_2 > p$. Cook's distance will be small if $\tilde{\delta} \approx 0$, $\hat{\beta}^{(2)} \approx \beta^{(1)}$, and $\underset{\sim}{\bar{X}}_2 \approx \mu_x^{(1)}$ while Bingham's statistic $T_1^2$ will be negligible if $\tilde{\delta} \approx 0$, $\underset{\sim}{\hat{\beta}}^{(2)} \approx \underset{\sim}{\hat{\beta}}^{(1)}$, and $\underset{\sim}{y}_2 \approx \underset{\sim}{\hat{y}}_{2(1)}$. Bingham's statistic has a component measuring the scatter of the deleted subset, which is absent in Cook's. Cook's statistic weights $\tilde{\delta}$ by a function of the distance between $\underset{\sim}{\bar{X}}_2$ and $\underset{\sim}{\mu}_x^{(1)}$; and each statistic measures the difference in $\hat{\beta}^{(2)}$ and $\underset{\sim}{\hat{\beta}}^{(1)}$ relative to a different metric. Hence, asymptotically, the difference in predictive mean vectors does not depend on the scatter of

deleted subsets about their regression lines, and the difference in predictive covariance matrices is independent of the distance between $\bar{\underset{\sim}{X}}_2$ and $\underset{\sim}{\mu}_x^{(1)}$. It is possible to determine a subset with zero influence by letting $\tilde{\delta} = 0$, $\hat{\underset{\sim}{\beta}}^{(2)} = \underset{\sim}{\beta}^{(1)}$ and $(1 - \frac{p}{n_2}) \frac{s_2}{\theta} = 1$. Such a subset would have the same regression line as the non-deleted subset, and hence point predictions would be the same. This subset would have moderate to large scatter depending on the magnitude of $\frac{p}{n_2}$. A very influential subset would have, relative to $\theta$, large $\tilde{\delta}$, large measures of the distance between $\hat{\underset{\sim}{\beta}}^{(2)}$ and $\underset{\sim}{\beta}^{(1)}$, large $s_2^2$, and a large measure of the distance between $\bar{\underset{\sim}{X}}_2$ and $\underset{\sim}{\mu}_x^{(1)}$; i.e. a very influential subset would be centered off the non subset-deleted regression line, would be alligned perpendicular to this line, would have large scatter about it's own line, and would be observed away from the center of the non subset-deleted data. These results are consistent with those in Johnson and Geisser (1979) except that here, predictive influence does not depend on some particular set of values at which one is assumed to be predicting.

Consider the expression (2.4.15) and note that, when $n_2 = 1$, the asymptotic predictive influence is not monotone and increasing in $T_i^2$ as was the case in Johnson and Geisser (1979). Note that $\frac{\partial}{\partial T_i^2}[\tilde{I}(f, f_{(i)})] \simeq \frac{1}{2}[1 + (\underset{\sim}{x}_i^* - \bar{\underset{\sim}{X}}_1) S_x^{(1)^{-1}} (\underset{\sim}{x}_i^* - \bar{\underset{\sim}{X}}_1)'] + T_i^2 - 1 > 0$ if and only if $T_i^2 + \frac{1}{2}(\underset{\sim}{x}_i^* - \bar{\underset{\sim}{X}}_1) S_x^{(1)^{-1}} (\underset{\sim}{x}_i^* - \bar{\underset{\sim}{X}}_1)' > \frac{1}{2}$. Hence $\tilde{I}(f, f_{(i)})$ is convex in $T_i^2$ unless $(\underset{\sim}{x}_i^* - \bar{\underset{\sim}{X}}_1) S_x^{(1)^{-1}} (\underset{\sim}{x}_i^* - \bar{\underset{\sim}{X}}_1)' > 1/2$, in which case $\tilde{I}(f, f_{(i)})$ is monotone and increasing in $T_i^2$. Note that $\min_{T_i^2} \tilde{I}(f, f_{(i)})$ is achieved for $T_i^2 \leq \frac{1}{2}$. The most influential single observation will, as usual, have a large value for $T_i^2$, and a large distance in the above sense, between $\underset{\sim}{x}_i^*$ and $\bar{\underset{\sim}{X}}_1$. The fact that

$\tilde{I}(f, f_{(i)})$ may be convex in $T_i^2$ merely reflects the fact that the first component is minimized for $T_i^2 = 0$, and the second component is minimized for $T_i^2 = 1$, and since the first component is now divided by 2, relatively less importance is attached here. Note that in Johnson and Geisser (1979) the asymptotic predictive influence was a monotonically increasing function of $T_i^2$, since the first component of P.I. was not divided by 2.

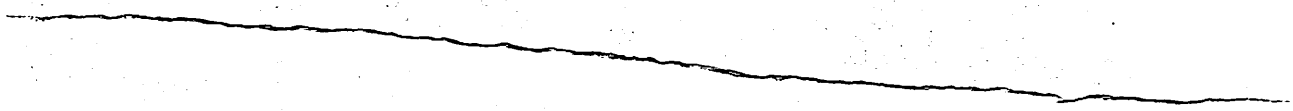
## 3. AN EXAMPLE

### (3.1) Introduction

We present an analysis of a data set which has already been discussed in great detail by Cook and Weisberg (1979). We consider a set of data taken from the 1975 Florida Area Cumulus Experiment (FACE) which was conducted to determine the merits of using silver iodide to produce rainfall increases, and to isolate some factors contributing to treatment unit additivity (Woodley et al., 1977). The target area consisted of about 3,000 square miles to the North and East of Coral Gables, Florida. In this experiment, 24 days in the summer of 1975 were judged suitable for seeding based on a suitability criterion, S. (For details see Woodley et al., 1977 or Cook and Weisberg, 1979). On each suitable day, the decision to seed was based on unrestricted randomization. The following variables were measured on each suitable day:

Echo Coverage (C) ——— Percent cloud cover in the experimental area, measured using radar in Coral Gables, Florida.

(P) ——— Total rainfall in the target area one hour before seeding (in $(\text{meters})^3 \times 10^7$).



Echo Motion (E) ——— A classification indicating a moving radar echo (1) or a stationary radar echo (2).

Response Variable (Y) —— The amount of rainfall that fell in the target area for a six-hour period on each suitable day (in $(\text{meters})^3 \times 10^7$).

The data is presented by Woodley et al., (1977) are reproduced in Table I. The variable

Time Trend ——— Number of days after the first day of the experiment (June 16, 1975 = 0).

is included, as in Cook and Weisberg (1979).

In addition to the suitability criterion, an attempt was made to use only days with $C \leq 13$ percent. Days with $C > 13$ percent were defined as disturbed days. From Table I we see that days 1 and 2 are disturbed. It is clear that day 2 is significantly disturbed; which will certainly result in large values for $tr(V_2)$ and $|I + V_2(I - V_2)^{-1}|$, and will cause weightings for residual differences to be heavy. It is to be expected that this observation will be included in the most influential subsets. This is shown to be the case for $n_2 = 1, 2, 3$. Cook and Weisberg (1979) have noted this fact and have chosen not to include this observation in their initial analyses since "the process under study may differ under the conditions of case 2." We prefer to analyze the full data set, and to then delete observation 2 and re-analyze in its absence.

We adopt the same initial model as Cook and Weisberg (1979):

$$(3.1.1)\quad L(Y) = \beta_0 + A\beta_1 + T\beta_2 + S\beta_3 + C\beta_4 + L(P)\beta_5 + E\beta_6 + (A \times S)\beta_{13} + (A \times C)\beta_{14} + (A \times L(P))\beta_{15} + (A \times E)\beta_{16}$$

Where $L(Y) = \log_{10}(Y)$ , $L(P) = \log_{10}(P)$.

See Cook and Weisberg (1979) for an explanation of the model.

Table I

Measurements from FACE, 1975

| CASE | A | T | S | C | P | E | SA | CA | PA | EA | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1.75 | 13.40 | .274 | 2 | 0 | 0 | 0 | 0 | 12.85 |
| 2 | 1 | 1 | 2.70 | 37.90 | 1.267 | 1 | 2.70 | 37.90 | 1.267 | 1 | 5.52 |
| 3 | 1 | 3 | 4.10 | 3.90 | .198 | 2 | 4.10 | 3.90 | .198 | 2 | 6.29 |
| 4 | 0 | 4 | 2.35 | 5.30 | .526 | 1 | 0 | 0 | 0 | 0 | 6.11 |
| 5 | 1 | 6 | 4.25 | 7.10 | .250 | 1 | 4.25 | 7.10 | .250 | 1 | 2.45 |
| 6 | 0 | 9 | 1.60 | 6.90 | .018 | 2 | 0 | 0 | 0 | 0 | 3.61 |
| 7 | 0 | 18 | 1.30 | 4.60 | .307 | 1 | 0 | 0 | 0 | 0 | .47 |
| 8 | 0 | 25 | 3.35 | 4.90 | .194 | 1 | 0 | 0 | 0 | 0 | 4.56 |
| 9 | 0 | 27 | 2.85 | 12.10 | .751 | 1 | 0 | 0 | 0 | 0 | 6.35 |
| 10 | 1 | 28 | 2.20 | 5.20 | .084 | 1 | 2.20 | 5.20 | .084 | 1 | 5.06 |
| 11 | 1 | 29 | 4.40 | 4.10 | .236 | 1 | 4.40 | 4.10 | .236 | 1 | 2.76 |
| 12 | 1 | 32 | 3.10 | 2.80 | .214 | 1 | 3.10 | 2.80 | .214 | 1 | 4.05 |
| 13 | 0 | 33 | 3.95 | 6.80 | .796 | 1 | 0 | 0 | 0 | 0 | 5.74 |
| 14 | 1 | 35 | 2.90 | 3.00 | .124 | 1 | 2.90 | 3.00 | .124 | 1 | 4.84 |
| 15 | 1 | 38 | 2.05 | 7.00 | .144 | 1 | 2.05 | 7.00 | .144 | 1 | 11.86 |
| 16 | 0 | 39 | 4.00 | 11.30 | .398 | 1 | 0 | 0 | 0 | 0 | 4.45 |
| 17 | 0 | 53 | 3.35 | 4.20 | .237 | 2 | 0 | 0 | 0 | 0 | 3.66 |
| 18 | 1 | 55 | 3.70 | 3.30 | .960 | 1 | 3.70 | 3.30 | .960 | 1 | 4.22 |
| 19 | 0 | 56 | 3.80 | 2.20 | .230 | 1 | 0 | 0 | 0 | 0 | 1.16 |
| 20 | 1 | 59 | 3.40 | 6.50 | .142 | 2 | 3.40 | 6.50 | .142 | 2 | 5.45 |
| 21 | 1 | 65 | 3.15 | 3.10 | .073 | 1 | 3.15 | 3.10 | .073 | 1 | 2.02 |
| 22 | 0 | 68 | 3.15 | 2.60 | .136 | 1 | 0 | 0 | 0 | 0 | .82 |
| 23 | 1 | 82 | 4.01 | 8.30 | .123 | 1 | 4.01 | 8.30 | .123 | 1 | 1.09 |
| 24 | 0 | 83 | 4.65 | 7.40 | .168 | 1 | 0 | 0 | 0 | 0 | .28 |

A = Action (0 = not seeded, 1 = seeded)

T = Time in days (June 16, 1975 = 0)

S = Seeding Suitability Criterion

C = Echo Coverage in Percent

P = Prewetness (in cubic meters x $10^7$)

E = Echo Motion (1 = moving radar echo, 2 = stationary radar echo)

SA = S x A

CA = C x A

PA = P x A

EA = E x A

Y = Rainfall (in cubic meters x $10^7$)

Our goal is to determine those subsets which are the most influential for the purpose of predicting future vectors at $X$, when the model (3.1.1) is given and fixed. This is a secondary goal for Cook and Weisberg (1979); their primary goal is to describe the difference $\Delta L(Y)$, between predicted rainfall for seeded and unseeded days. They are further interested in determining the "most appropriate" model for achieving this purpose, in conjunction with their outlier analysis. We will not pursue this issue.

For the cases $n_2 = 1, 2, 3$, and the model (3.1.1), we compute the approximate P.I.F. $(n)^2 \cdot \hat{I}(f_{(2)}, f)$, which is defined in (2.4.13). We consider the first and second components separately, and the sum. Consequently, subsets which affect mean vectors may be distinguished from those which affect covariance structures. A summary of these results for the full data set, as well as the full data set minus case 2, are given in Table II.

(3.2) Analysis of the Data

A computer program was written to compute the statistics defined in (2.4.13) for all $\binom{n}{n_2}$ subsets; $n_2 = 1, 2, 3$, $n = 23, 24$. Subsets were ordered according to the magnitudes of $\hat{I}(f_{2)}, f)$. Computer output includes the first and second components of $(n)^2\ \hat{I}(f_{(2)}, f)$ as well as the sum. An obvious problem is the magnitude of the number of calculations necessary for even moderate $n_2$. When $n = 24$ and $n_2 = 3$, 2024 calculations of $\hat{I}$ were necessary. Cook and Weisberg (1979) have determined some inequalities which make it unnecessary to do all calculations. Such an approach is indicated here.

Table II

Top 5 most influential subsets $n_2 = 1, 2, 3$.

| | Full Data Set | | | | Observation 2 Deleted | | | |
|---|---|---|---|---|---|---|---|---|
| | Observation | First Component | Second Component | Sum | Observation | First Component | Second Component | Sum |
| $n_2 = 1$ | 2 | 7.02 | 16.32 | 23.34 | 7 | 3.71 | 5.13 | 8.84 |
| | 7 | 4.01 | 5.26 | 9.27 | 24 | 2.21 | 3.79 | 6.00 |
| | 24 | 2.39 | 3.92 | 6.31 | 6 | 2.25 | .81 | 3.06 |
| | 6 | 2.53 | .79 | 3.32 | 18 | 1.18 | 1.68 | 2.86 |
| | 18 | 1.43 | 1.60 | 3.03 | 17 | .56 | .93 | 1.49 |
| $n_2 = 2$ | 2, 5 | 40.10 | 22.85 | 62.95 | 7, 24 | 2.09 | 14.81 | 16.90 |
| | 2, 21 | 18.08 | 29.09 | 47.17 | 4, 24 | 4.22 | 8.75 | 12.97 |
| | 2, 15 | 1.35 | 30.30 | 31.65 | 1, 17 | 8.17 | 4.18 | 12.35 |
| | 2, 14 | 8.21 | 21.24 | 29.45 | 7, 16 | 5.23 | 6.22 | 11.45 |
| | 2, 11 | 9.13 | 19.12 | 28.25 | 4, 7 | 6.35 | 4.86 | 11.21 |
| $n_2 = 3$ | 2,5,21 | 125.22 | 41.87 | 167.09 | 5,11,23 | 83.43 | 49.02 | 132.45 |
| | 2,5,23 | 49.33 | 56.88 | 106.21 | 1,13,17 | 22.39 | 13.77 | 36.16 |
| | 2,14,21 | 32.91 | 47.29 | 80.20 | 1,6,13 | 18.17 | 12.78 | 30.95 |
| | 2,5,15 | 32.56 | 44.56 | 77.12 | 4,16,24 | 9.61 | 18.16 | 27.77 |
| | 2,5,14 | 45.77 | 29.31 | 75.08 | 1,8,17 | 19.28 | 3.59 | 22.87 |

We see from Table II that observations 2, 7, and 24 are the most influential when $n_2 = 1$. Note that if observations were ordered according to the first component, $2 \succ 7 \succ 6 \succ 24 \succ 18$; and when ordered by the second component, $2 \succ 7 \succ 24 \succ 18 \succ 6$. Judging from the magnitudes of each of the terms, it seems clear that case 2 is very influential for both point and ellipsoidal predictions, and that it is significantly more influential when considering the latter. A criticism which might be leveled at this point, is that no method has been proposed for determining the significance of magnitude of the P.I.F. . A possibility will be suggested in Section (3.3). Still, it is clear that case 2 is an outlier and that

it significantly affects predictive inference based on this data set.

Looking at the $n_2 = 2$ and $n_2 = 3$ cases, it is clear from Table II that case 2 will be included in most influential subsets. We find that the most influential outlier pair is (2, 5) due to a large second component, and even larger first component. Again, subsets would be ordered differently according to first and second components. When $n_2 = 3$, the most influential triple is (2, 5, 21). Here again, the first component is affected the most. Rather than dwell on the possibilities here, we choose to delete case 2 and perform a more careful analysis then. It is glaringly apparent that case 2 will affect our analysis and interpretation to a large extent, and it seems that its inclusion will mask salient features of the remaining data set.

We see from Table II that observations 7, 24, and 6 are most influential when case 2 is deleted, and that (7, 24) is the most influential pair. Note that this pair is not apparently influential due to its effect on mean vectors, but due to its effect on covariance structures. This is consistent with the results of Cook and Weisberg (1979). They note that $D_s^2$ for this case is not large (.455), but that $T_1^2$ is significant at the .064 level, using a Bonferroni inequality. Our first component is essentially $D_s^2$, and our second component depends heavily on the magnitude of $T_1^2$. After further analysis, Cook and Weisberg (1979) choose to delete (7, 24) as well as case 2. They note that 7, and 24 seem to be singly influential as well as pairwise influential and that both observations came on days which were not seeded, and where unusually low responses were observed. Their conclusion is that the pair (7, 24) does not belong to the assumed model.

We consider the $n_2 = 3$ case and note from Table II that the triple

(5, 11, 23) is very influential with respect to both components; especially the first. At first glance, this seems surprising. Observation 24 shows up in the 4th most influential subset, and the pair (7, 24) shows up in the 7th most influential subset along with case 8. It is very interesting to note that the first seven most influential pairs involve unseeded days, and the first 10 most influential triples either involve all unseeded or all seeded days. The triple (5, 11, 23) involves only seeded days where responses were relatively low. The most influential pair from (5, 11, 23) is (5, 23) which is 85th in order of magnitude. It is apparent that some combination of the orientation, scatter, and location of center of these points, relative to the rest, makes them quite unique as a unit. Observe that both components are very large, but that the first is appreciably larger than the second. Consideration of (2.3.26) and (2.4.11) implies that this subset may be distantly centered from the non-deleted set; i.e. that $\underset{\sim}{X}_2$ is distant from $\underset{\sim}{X}_1$. This, in conjunction with the fact that these days were all seeded days with relatively low responses, indicates that deletion of these data points, tends to void the subspace which they span, and that this subspace is distant from that spanned by the rest of the data. This view is supported by the fact that these observations are neither singly nor pairwise influential, and also by a visual scan of the data.

As a final comment, note that

$$\bar{V}_2 = \text{Var}_{\text{samp.}}\ (n_2^{-1} \sum_{j=1}^{n_2} \hat{Y}_{2j}) = n_2^{-2} \underset{\sim}{e}'_{n_2} V_2 \underset{\sim}{e}_{n_2} = n^{-1} \{1 + (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}}) S_x^{-1} (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}})'\}$$

by an application of (2.3.8). Further note from (2.3.26) that

$$D_S^2 = \frac{n_2^2}{n} [\hat{\delta}\{1 + (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}}) S_x^{-1} (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}})'\} + (\underset{\sim}{s}_{yx} - \hat{\underset{\sim}{\beta}}^{(1)\prime} S_x^{(2)}) S_x^{-1} (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}})']^2 / \{1 + (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}}) S_x^{-1} (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}})'\}$$

$$+ (s_{yx}^{(2)} - \beta_x^{(1)\prime} S_x^{(2)}) [S_x + (\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}})'(\underset{\sim}{\bar{X}}_2 - \underset{\sim}{\bar{X}})]^{-1} (s_{xy}^{(2)} - S_x^{(2)} \hat{\underset{\sim}{\beta}}^{(1)}) .$$

Now if $s_{xy}^{(2)} \approx S_x^{(2)}\hat{\underset{\sim}{\beta}}^{(1)}$, it follows that Cook's distance is

$$\frac{D_S^2}{s \cdot p^2} \approx \frac{n_2^2}{p} \frac{\hat{\delta}^2}{s^2} \bar{V}_2 ,$$

which is just $n_2^2 D_i$ for $Y_i = \bar{Y}_2$, $\underset{\sim}{X}_i = (1, \bar{\underset{\sim}{X}}_2)$.

Hence, in this case, a subset of size $n_2$ is $n_2^2$ times as influential as a single observation which is observed at the center of the deleted subset. At any rate, Cook's distance will be large when $(\bar{Y}_2, \bar{\underset{\sim}{X}}_2)$ is distant from $(\bar{Y}, \bar{\underset{\sim}{X}})$ and the sign of $(\underset{\sim}{s}_{yx}^{(2)} - \hat{\underset{\sim}{\beta}}^{(1)\prime} S_x^{(2)}) \; S_x^{-1}(\bar{\underset{\sim}{X}}_2 - \bar{\underset{\sim}{X}})'$ is positive. This seems to be the case for our example.

(3.3) Conclusions

Cook and Weisberg (1979) concluded that observations (2, 7, 24) should be removed from the data due to their lack of conformity to the assumed model. We find no evidence which conflicts with this view. Observation (2, 7, 24) are singly influential and in the absence of observation 2, observation (7, 24) are both singly influential and pairwise influential. While the observations correspond to unusually low responses on unseeded days, their deletion does not appear to void an important subspace of the space of independent variables spanned by the full data set. We finally note that the triple (5, 11, 23), is extremely influential in the absence of observation 2, and is influential to a large degree due to the fact that these points do span an important subspace of the space spanned by the full data set.

It would be useful to be able to determine statistically when max $I(f_{(2)}, f)$ is large. This might be accomplished by determining the bootstrap distribution, Effron (1979), of max $I(f_{(2)}, f)$ when the most

influential subset has been removed. The observed $\max\ I(f_{(2)}, f)$ could then be compared to percentage points of the bootstrap distribution.

## 4. Appendices

### (4.1) Proof of Propositions (2.3.3)-(2.3.5)

Proof of Proposition (2.3.3):

The result (2.3.15) is obvious since M is idempotent of rank p.

Now define $U_{ij} = X_i S_1^{-1} X_j'$, $V_{ij} = X_i S^{-1} X_j'$ $\quad i,j = 1, 2.$

Then by some algebra and Frobenius' theorem,

$$|I + M_{(2)}| = \begin{vmatrix} I + U_1 & U_{12} \\ U_{21} & I + U_2 \end{vmatrix} = |I + U_1||I + U_2 - U_{21}(I+U_1)^{-1}U_{12}|.$$

But $U_1$ is idempotent of rank p and $(I+U_1)^{-1} = I - \frac{1}{2} U_1$. Hence by (2.3.3) and (2.3.4) the above equals

$$2^p|I + U_2 - U_{21}(I - \tfrac{1}{2}U_1)U_{12}| = 2^p|I + \tfrac{1}{2}U_2| = 2^p|I + \tfrac{1}{2}V_2(I-V_2)^{-1}|$$

and (2.3.16) is true. ■

Proof of Proposition (2.3.4):

It is implicit in Geisser (1965) that

$$(I + M_{(2)})^{-1} = I - X(S + S_1)^{-1}X' .$$

To see this result, observe from (2.3.1) that

$$I - X(S + S_1)^{-1}X' = I - M_{(2)}(I - (I + M_{(2)})^{-1}M_{(2)}$$

$$= I - M_{(2)}(I + M_{(2)})^{-1} = (I + M_{(2)})^{-1} .$$

Further, by (2.3.2) and some algebra

$$I - X(S + S_1)^{-1}X' = I - X(2S - S_2)^{-1}X'$$

$$= I - X(\tfrac{1}{2}S^{-1} + \tfrac{1}{4}S^{-1}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2S^{-1})X'$$

$$= I - \tfrac{1}{2}M - \tfrac{1}{4}XS^{-1}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2S^{-1}X'$$

and the result obtains. ■

Proof of Proposition (2.3.5):

Since $M$ is idempotent and $(I + M)^{-1} = I - \tfrac{1}{2}M$, it follows that

$$\operatorname{tr}(I + M)^{-1}(I + M_{(2)}) = \operatorname{tr}(I - \tfrac{1}{2}M + \tfrac{1}{2}M_{(2)})$$

$$= n - \tfrac{1}{2}p + \tfrac{1}{2}\operatorname{tr}\begin{pmatrix} U_1 & U_{12} \\ U_{21} & U_2 \end{pmatrix} = n - \tfrac{1}{2}p + \tfrac{1}{2}\operatorname{tr}(U_1 + U_2) .$$

But $U_1$ is also idempotent of rank $p$ and $U_2 = V_2(I - V_2)^{-1}$ hence the above equals

$$n + \tfrac{1}{2}\operatorname{tr} V_2(I - V_2)^{-1}$$

and (2.3.18) obtains. Now by (2.3.17) and the fact that $M$ is idempotent

$$\operatorname{tr}(I + M_{(2)})^{-1}(I + M) =$$

$$\operatorname{tr}[(I - \tfrac{1}{2}M - \tfrac{1}{2}XS^{-1}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2S^{-1}X')(I + M)]$$

$$= \operatorname{tr}[I - \tfrac{1}{2}XS^{-1}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2S^{-1}X']$$

$$= n - \tfrac{1}{2}\operatorname{tr}[V_{12}(I - \tfrac{1}{2}V_2)^{-1}V_{21} + V_2(I - \tfrac{1}{2}V_2)^{-1}V_2].$$

But the trace operator is cyclic, so the above equals

$$n - \tfrac{1}{2}\operatorname{tr}[S^{-1}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2S^{-1}S_1 + V_2(I - \tfrac{1}{2}V_2)^{-1}V_2]$$

$$= n - \tfrac{1}{2}\operatorname{tr}[S^{-1}X_2'(I - \tfrac{1}{2}V_2)^{-1}X_2S^{-1}(S - X_2'X_2) + V_2(I - \tfrac{1}{2}V_2)^{-1}V_2]$$

$$= n - \tfrac{1}{2}\operatorname{tr}[V_2(I - \tfrac{1}{2}V_2)^{-1}]$$

and the result (2.3.19) obtains. ■

(4.2) Proof of the Results (2.3.29)

To derive the first part of (2.3.29), it is enough by (2.3.21), (2.3.26) and

convergence assumptions to show that

$$(4.2.1)\quad \frac{1}{2}\,\mathrm{tr}\, V_2(I - V_2)^{-1} - \ln\left|I + \frac{1}{2} V_2(I - V_2)^{-1}\right| = o(n^{-1}) .$$

Let $\{o_{ij}\}_k$ denote a $k\times k$ matrix such that $o_{ij} = o(1)$; $i, j = 1, \ldots, k$. Then since $V_2(I-V_2)^{-1} = U_2 = \{o_{ij}\}_{n_2}$ by (2.3.3), (2.3.4), (2.3.7) and convergence assumptions, it follows that

$$\frac{n}{2}\,\mathrm{tr}\, V_2(I - V_2)^{-1} = \frac{n}{2} \sum_{j=1}^{n_2} o_{jj} .$$

Further, by the definition of the determinant, two Taylor expansions, and the above

$$(4.2.2)\quad n \ln\left|I + \frac{1}{2} V_2(I-V_2)^{-1}\right| = n \ln\left|I + \frac{1}{2}\{o_{ij}\}_{n_2}\right|$$

$$= n \ln\left\{\left[\prod_{j=1}^{n_2} (1 + \frac{1}{2} o_{jj})\right] + o(n^{-1})\right\}$$

$$= n \ln\left\{\prod_{j=1}^{n_2} (1 + \frac{1}{2} o_{jj})\right\} + o(1)$$

$$= n \sum_{j=1}^{n_2} \frac{1}{2} o_{jj} + o(1) .$$

Hence,

$$\frac{n}{2}\,\mathrm{tr}\, V_2(I-V_2)^{-1} - \ln\left|I + \frac{1}{2} V_2(I - V_2)^{-1}\right| = o(1)$$

and (4.2.1) is true, and hence, the second part of (2.3.29) is true.

Recall that $(I - \frac{1}{2}V_2)^{-1} = \sum_{k=0}^{\infty} (\frac{1}{2})^k V_2^k$ . Then since $V_2 = \{o_{ij}\}_{n_2}$ ,

$$(4.2.3)\quad (I - \frac{1}{2} V_2)^{-1} = I + \{o_{ij}\}_{n_2},$$

hence

$$\theta Q_2 = S - \frac{1}{2} X_2' (I - \frac{1}{2} V_2)^{-1} X_2 = S - \frac{1}{2} X_2' (I + \{o_{ij}\}_{n_2}) X_2 .$$

Now since $\hat{\beta}_{(2)} \overset{a.s.}{\to} \beta$ as $n \to \infty$,

$$(y_2 - \hat{y}_{2(2)})' X_2 S^{-1} X_2' (I - \tfrac{1}{2} V_2)^{-1} X_2 S^{-1} X_2 (y_2 - \hat{y}_{2(2)})$$

$$= (y_2 - \hat{y}_{2(2)})' \{o_{ij}\}_{n_2} (I + \{o_{ij}\}_{n_2}) \{o_{ij}\}_{n_2} (y_2 - \hat{y}_{2(2)}) = o(n^{-1}) \text{ a.s.} .$$

Hence by (2.3.12) and the above

$$\theta D^2_{Q_2} = D^2_S + o(n^{-1}) \qquad \text{a.s.}$$

and so

$$nD^2_{Q_1} \approx nD^2_{Q_2} \approx \tfrac{n}{2} D^2_{Q_3} \qquad \text{a.s.} .$$

It remains to show that

$$(4.2.4) \quad \ln|I + \tfrac{1}{2} V_2 (I - V_2)^{-1}| - \tfrac{1}{2} \text{tr } V_2 (I - \tfrac{1}{2} V_2)^{-1} = o(n^{-1}) .$$

But by (4.2.3)

$$\tfrac{1}{2} \text{tr } V_2 (I - \tfrac{1}{2} V_2)^{-1} = \tfrac{1}{2} \text{tr } [V_2 + V_2 \{o_{ij}\}_{n_2}]$$

$$= \tfrac{1}{2} \{\text{tr}[\{o_{ij}\}_{n_2}] + o(n^{-1})\} = \tfrac{1}{2} \sum_{j=1}^{n_2} o_{jj} + o(n^{-1}) .$$

Hence, (4.2.4) follows from (4.2.2); and (2.3.29) is true.

## REFERENCES


Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis. Cambridge, Cambridge University Press.

Behnken, D. W. and Draper, N.R. (1972). "Residuals and their variance patterns," Technometrics 14, 102-111.

Bingham, Christopher (1977). "Some identities useful in the analysis of residuals from linear regression." Unpublished Technical Report No. 300. School of Statistics, University of Minnesota.

Cook, R.D. (1977). "Detection of influential observations in linear regression." Technometrics 19, 15-18.

Cook, R.D. (1979). "Influential observations in linear regression." JASA 74, 169-174.

Cook, R.D. and Weisberg, S. (1979). "Finding influential cases in regression: a review." Unpublished Technical Report No. 338. School of Statistics, University of Minnesota.

Efron, B. (1979). "Bootstrap methods: another look at the jackknife." Annals of Statistics 7, 1-26.

Geisser, S. (1965). "Bayesian estimation in multivariate anlysis." Ann. Math. Stat. 36, 150-159.

Geisser, S. (1971). "The inferential use of predictive distributions." Foundations of Statistical Inference. Edited by Godambe & Sprott. Toronto, Holt, Rinehart and Winston, 456-466.

Johnson, W., & Geisser, S.(1979) "Assesing the predictive influence of observation," Unpublished Technical Report No. 355, School of Statistics, University of Minnesota.

Kullback, S. (1968). Information Theory and Statistics. Gloucester, Mass., Peter Smith.

Kullback, S. and Leibler, R.A. (1951). "On information and sufficiency." Annals of Math. Stat. 22, 79-86.

Murray, Gordon D. (1977). "A note on the estimation of probability density functions." Biometrika 64, 150-152.

Woodley, W.L. et al. (1977). "Rainfall results, 1970-75: Florida area cumulus experiment." Science 195 (25 Feb. 1977), 735-742.